Final Report to

Nebraska Department of Education

Examining the Potential for Selected NRTs and Locally Developed CRTs to Classify Students into Performance Categories in Reading and Mathematics

Prepared by

James C. Impara, Ph.D.

Chad W. Buckendahl, Ph.D.

Abdullah Ferdous, M.A.

Renee Jacobson, Ed. S.

The Buros Institute for Assessment Consultation and Outreach is

A Division of the

Oscar and Luella Buros Center for Testing

University of Nebraska - Lincoln

May 2004

Examining the Potential for Selected NRTs and Locally Developed CRTs to Classify Students into Performance Categories in Reading and Mathematics

EXECUTIVE SUMMARY

This project had two phases and was undertaken during the period from March 2003 to April 2004. The data for this study were collected in two phases. Phase 1 was the determination of performance levels in reading and mathematics. This took place in July 2003 and a report was delivered in September 2003. This report covers the second phase of the project in which data were collected to determine the extent that 3 NRTs from different publishers and 21 CRTs (11 in reading and 10 in mathematics) could be used to classify students as performing at Advance, Proficient, Progressing, or Beginning levels on selected standards. In addition to this main focus (the determination of the sufficiency of the assessments to make performance level classifications) there was a second study that looked at the consistency of teachers across multiple sites to make the same performance level judgments about a sample of assessments.

A total of 66 teachers met at one of three locations around the state to judge the assessments in reading. An additional 52 teachers met at these same locations to judge the mathematics assessments. After an introduction to the study and undertaking a practice activity, teachers spent approximately a day evaluating the tests at their grade level.

The results of this activity suggest that the NRTs have some utility for classifying students as either Proficient or Below Proficient on the one reading standard that was examined. Some additional, more precise classification is also possible, but such classifications should be made with caution because, in some cases multiple subtests are involved and the proportion of items that focus on the specific standard may be a relatively small proportion of the items on the subtest. This is even truer in mathematics where the utility of classifying student on two standards was examined. On the standard related to Computation and Estimation, the NRTs were comparable to the utility for reading. Similarly, the items were often across more than one subtest, so the proportion of the total items that focus on the standard is not known. However, for the standard related to Data Analysis, Probability and Statistical Concepts, there tended to be fewer items, all in one subtest and fewer opportunities to make performance level decisions.

The CRTs that were examined included some assessments that were locally developed and used only in the local district. Other assessments had been developed in consortia and were used in multiple districts. In general the utility of these assessments to classify students into multiple performance levels is mixed. At grade 4 in both reading and mathematics the assessments of about 67% of the standards provided classification into either Proficient or Below Proficient categories. This percentage declines rapidly to about 50% at grade 8 and about 25% at high school. If more precise classifications are desired (e.g., Advanced, Proficient, Progressing, Beginning) these percentages drop dramatically. The worst case is in attempting to classify students as Advanced. Across all grade levels, assessments and standards in reading (3 x 11 x 6) there were only 13 instances when a rating of Advanced could be made. The situation is only slightly better in mathematics where there are 22 such instances.

The study related to consistency of judgments across settings, suggested that some caution should be undertaken in interpreting the results related to the CRTs. Although it was often the case that there was agreement between two of the three sets of teachers, and occasional agreement among all three sets, there were also some assessments on which there was virtually no agreement. In short, there was no consistent pattern of agreement. That is, for some assessments the teachers in the Eastern and Central regions agreed with each other on their item classification judgments, for other assessments, the Western and Central teachers agreed, and for yet other assessments the Eastern and Western teachers agreed. These were not systematic for the same standards.

A number of recommendations were made. Some of these recommendations focus on the process (e.g., making it clearer how the performance level definitions should be used). Other recommendations focus on the results, particularly on the utility of the assessments to be used for making performance level classifications. All of the recommendations are listed below.

Recommendations

Process

- The notion of overlapping participants may be sound, but the time span between the development of the definitions and the operational study needs to be much closer in time so that the participants who develop the definitions do not have time to forget the discussions and rationale for the decisions made at the initial meeting.
- Before undertaking further studies using performance level definitions, the definitions should be reviewed and ambiguities and inconsistencies eliminated.
- A rule of precedence must be developed when more than one standard might be applicable (either directly or indirectly) to an assessment task.
- Clarify the weight of the rubric and the weight of the assessment task when there appears to be a conflict. This may require expanding the performance level definitions to include not only content, but also skills that are needed to demonstrate content knowledge (i.e., how the content is assessed).
- Continually emphasize to the teachers that the decision about the performance level of an assessment task should not be based on their students but instead is based on the performance level definitions provided and on the rules of precedence and weights developed based on the above recommendations.
- In future studies, a better explanation of the outcomes of the process and reassurance that the results will be confidential (to other districts and the NDE) along with assurances that the results will be shared with the district, may help with getting cooperation.
- Districts should be advised that more direction to teachers is needed. For example, rather than just leave open the selection of a reading passage, teachers should be provided a list of passages that are of similar difficulty.

- The NDE needs to work with districts to help them understand the difference between when rubrics are appropriate and when they are not. Moreover, the districts need to know that using the rubric to define the performance levels is not an appropriate procedure, because such a process bypasses the standard setting process and makes the performance level classification arbitrary.
- The NDE should encourage all districts to develop performance level definitions and to apply these definitions independent of the rubrics used to score assessment tasks. If this is not done, then the assessment tasks should be evaluated by the district to make sure that the assessment tasks and the score levels accurately reflect the performance level definitions.
- Practice tests should provide examples of all performance levels for all standards. They should be exemplary tests that represent quality assessments.

Assessment Utility

- Continue using the NRTs for classifying students as either Proficient or Below Proficient, but stop using NRTs for classifying students into four performance levels, especially using the arbitrary cut points of the 75%ile and 25%ile.
- Districts that are using these NRTs to make classification decisions should be doing so with extreme caution. It would be useful to supplement these tests with well-constructed CRTs to be more comfortable in the classification of students into performance levels. The items related to the strand 5 standard represent only a minority of the items on the subtest on which theses items are found, thus using the 50%ile (or any other point) as the dividing point for Proficient or Below Proficient will probably result in many misclassifications.
- Districts should carefully review their assessments to insure that the assessment tasks, and where appropriate the assessment rubrics, provide opportunities for students at all performance levels to demonstrate their knowledge and skill relative to all standards. Districts should start with their assessments at the high school level.
- The NDE should place little confidence in the districts' classifications of students based on these assessments. The reported data should be collapsed into only two categories Proficient and Below Proficient to obtain a more accurate representation of student performance levels.

Examining the Potential for Selected NRTs and Locally Developed CRTs to Classify Students into Performance Categories in Reading and Mathematics

INTRODUCTION

This report is the second and final report from this study. The study involved two phases. In phase one, performance level definitions were developed for Reading and Mathematics. Specifically, performance level definitions were developed for Reading at grades 4, 8 and high school and for six reading standards at each of these levels. In addition, performance level definitions were developed for mathematics overall (inclusive of all grades) and for each of the six mathematics strands used in Nebraska's content standards. In each content area four performance levels are defined: Beginning, Progressing, Proficient, and Advanced. For detailed information on these definitions see Impara, Buckendahl, and Jacobson (September, 2003) Phase One Report Examining the Potential for Selected NRTs and Locally Developed CRTs to Classify Students into Performance Categories in Reading and Mathematics: Developing Performance Level Definitions.

In phase two as detailed in this report, panelists used the performance level definitions developed in phase one to classify assessment items/tasks into their appropriate performance level. The purpose of phase two was to determine if a sample of commercially available norm-referenced tests (NRTs) and locally developed criterion-referenced tests (CRTs) would provide sufficient measurement information to permit classifying students into the four performance level categories.

The rationale for conducting this study is that one of the requirements of the No Child Left Behind legislation is that states report on the extent that students at selected grade levels are proficient in the state's standards. Moreover, although this reporting requirement employs a dichotomous classification of students (either proficient or not), other elements of the legislation require that students potentially be classified into one of several categories. Nebraska has named these categories Beginning, Progressing, Proficient, and Advanced. Students are to be classified on the basis of an assessment of skills and knowledge associated with state content standards in reading and mathematics.

The performance level definitions developed in phase one of this study will not be promulgated by the Nebraska Department of Education. These definitions were developed exclusively for use in phase two of this study. However, these performance level definitions are available to local districts to use as models for developing their own local definitions.

This report of the work in phase two will be sent to the Nebraska Department of Education and it will be shared with the publishers and districts that provided their assessment materials for the study.

The report is divided into several sections. The first section is a concise overview of the study. That section is followed by a description of the methodology used in the study, including some of the rationale for decisions that were made. The third section includes

specific results of the study[1]. Finally, section 4 provides conclusions and recommendations about the extent that these sample assessments can be used to classify students in multiple performance levels. Appendices show the results for each of the assessments evaluated in this study

METHODS AND PROCEDURES

In order to conduct this study a number decisions were made. These decisions included, which standards would be included, which norm-referenced tests would be included, performance level definitions for reading and mathematics, the methods and criteria for obtaining assessments from local districts, identifying teachers to participate, where and when teachers would come together to rate assessment tasks; methods for analyzing the teachers' ratings of the assessment tasks; and the structure of the reporting format. Each of these decision points is described below. Additionally, the structure of the meetings at which assessment task ratings were made is also described.

Selecting standards

This decision was made in phase one of the study and is described in detail in the phase one report. In summary, the principal investigators, accompanied by personnel from the NDE came together to decide on the standards. The standards that were identified by this group are listed below.

Reading:

Grade 4: 4.1.1, 4.1.2, 4.1.3*, 4.1.4, 4.1.5, 4.1.6

Grade 8: 8.1.1*, 8.1.2, 8.1.3, 8.1.4, 8.1.5, 8.1.6

Grade 12: 12.1.1*, 12.1.2, 12.1.3, 12.1.4, 12.1.5, 12.1.6

Mathematics:

Grade 4: 4.1.3, 4.2.1*, 4.3.4, 4.4.2, 4.5.1*, 4.6.2;

Grade 8: 8.1.4, 8.2.2*, 8.3.2, 8.4.1, 8.5.2*, 8.6.3;

Grade 12: 12.1.2, 12.2.1*, 12.3.1, 12.4.5, 12.5.1*, 12.6.3

Standards marked with an asterisk are measured by NRTs available from commercial publishers.

Selecting norm-referenced and district criterion-referenced assessments

The criteria for selection of norm-referenced assessments were 1) that the test was not going to be revised within the next year or two, thus the results would remain current for the near term and 2) that the tests were previously judged to assess the standards that were selected for inclusion in the study. This previous judgment was based on the most recent alignment student conducted by NDE in 2001 (see www.nde.state.ne.us/stars for the full report). Clearly, NRTs and standards were selected simultaneously. These criteria resulted in the selection of the Iowa Tests of Basic Skills (grades 4 & 8), the Iowa Tests

[1] No publishers or districts are identified other than in Table 1. All results are anonymous. The publishers and districts each received a report of the results of the study that provided the key to identifying their own results, but not the results of others.

of Educational Development (grade 11), the California Achievement Test – 6th edition (also referenced as Terra Nova II); and the Metropolitan Achievement Test – 10th edition.

The following criteria were used in making the determination of which districts would be eligible to be invited to provide their assessments for review. First, there are a representative number of district assessments from each region of the state, West, Central, and East (approximately 5 to 6 per region). Second, the assessments examined are "standardized" within the district. That is, all students in the district are assessed using the same assessment tasks for the standards being selected. Third, districts have assessment tasks for each of the selected standards (this is an issue only for Grade 12 as one of the selected strands and its related standards is currently voluntary). Fourth, that each selected district's assessment has been rated at least Very Good in terms of overall assessment quality. Fifth, that a range of district sizes be included. Specifically that there be some districts that have only one classroom at the selected grade level and there are some districts that are larger. A district could have provided assessments at all three grade levels if all three grade levels are represented in the district. Thus, for the three grade levels, a total of 9 - 12 district assessments for each of six standards were anticipated.

These criteria were not always met. Districts were hesitant to provide their assessments for review and obtaining cooperation proved to be very difficult. In total only 12 districts provided their assessments in Reading and 11 districts provided their assessments in Mathematics. The names of the districts are shown in Table 1. Some districts did not provide assessments for all grade levels (some did not have all grade levels, some did not want to share their assessments for some grade levels) and some did not have assessments that could be configured for rating[2]. Two districts' assessment materials were used for practice, so the operational ratings were done on 11 and 10 districts' assessment materials in reading and mathematics, respectively.

Because the number of districts that provided their assessments was fewer than had been planned, a modification in the study design was undertaken. This modification was a positive change. The modification entailed using some districts' assessments in multiple rating sessions (see below for a description of the different meeting locations). This change permitted an opportunity to investigate the extent that different teachers who had been trained similarly would rate the same materials in the same way. Thus, the districts

[2] The way some of the assessments were designed, it was not possible for them to be rated. For example, in reading, some assessment tasks required the teacher to identify a selection and have their students either answer general questions about the selection or do a report on the selection. Because the difficulty of the task depends almost entirely on the specific selection, it could not be rated. Similarly, some assessments specified a particular reading selection, but the selection was not provided (it may have been referenced in the assessment description), thus unless the teachers who were selected to participate in the rating task were familiar with the particular selection, no rating was possible. These situations were most prevalent in reading. In mathematics, some assessments were configured for rating, but were not rated because the participants indicated that these assessments were not properly aligned to the standard, thus rating was not possible.

are divided into those that were specific to their region and those that were rated by all of the study participants.

Table 1. Districts that volunteered their assessments for review in reading and mathematics.

Reading	Mathematics
East Elkhorn Johnson-Brock Nemaha Valley	East Lodgepole (this district also was rated in the West) Omaha Public Schools Nebraska City
Central Central City High Plains Fairbury	Central Eustis-Farnam Seward Valentine
West Alliance ESU 13 Maxwell	West Alliance ESU 13 Lodgepole (also rated in the East)
Common Raymond Central West Boyd Unified	Common Nebraska Unified Raymond Central
Practice Plattsmouth	Practice Beatrice

Meeting locations

To encourage participation, three locations for the operational classification of assessment tasks into performance categories, were selected: Lincoln, Kearney, and Scottsbluff. Thus, one meeting was held in each region of the state. It was the intention that in addition to one NRT being rated, assessments from that region would also be rated. Because of the lower than expected level of cooperation by school districts in sharing their assessments, some districts were rated in all three locations and one district's mathematics assessment was rated in two locations.

Selecting teachers

Because the study had two phases and because the meaning of the performance level definitions was a critical element of the rating of assessment tasks, teachers who had participated in phase one, development of performance level definitions provided a cadre of participants for phase two. Thus, two groups of teachers were recruited to participate in this study. Group 1 participated in both the determination of performance level definitions and the rating of assessment tasks, and Group 2 participated only in rating assessment tasks. Thus it was anticipated that up to six teachers (two at each grade level) from Group 1 would participate in each of the three regional meetings to rate assessment

tasks. The rationale for the Group 1 teachers was that they could provide insights into the development and meaning of the performance level definitions during the rating process. This, however, was not the case. The time frame for the study was such that most of the teachers who participated in the development of the definitions in late July 2003 did not recall the deliberations or outcomes in sufficient detail by the time they met again to make the ratings (either September, October, or November, 2003) to be helpful in the rating process.

As was the case in obtaining assessment materials for review, recruiting teachers was also a challenge. Not all of the Group 1 teachers elected to participate in the subsequent regional meetings to rate assessments. In addition, obtaining additional teachers often proved difficult. The number of teachers at each of the meeting sites is shown in Table 2.

Table 2. Number of teachers who participated in rating assessment tasks at each meeting.

Reading			Mathematics		
East	Grade	Teachers	East	Grade	Teachers
	4	9		4	9
	8	6		8	6
	H.S.	5		H.S.	6
Central	Grade	Teachers	Central	Grade	Teachers
	4	6		4	6
	8	4		8	5
	H.S.	7		H.S.	6
West	Grade	Teachers	West	Grade	Teachers
	4	7		4	4
	8	6		8	7
	H.S.	8		H.S.	5

Conducting the meetings

Six meetings were held in three locations. The meetings were in Lincoln on September 29-30 (Reading) and September 30-October 1 (Mathematics), in Kearney on November 3-4 (Reading) and November 4-5 (Mathematics), and in Scottsbluff on October 13-14 (Reading) and October 14-15 (Mathematics). The Reading meetings all started at 8:00 and ended in the morning of the following day. The Mathematics meetings all started at 1:00 and ended the afternoon of the next day. All six meetings were conducted in a similar fashion, so only one description is provided.

Prior to each of the meetings the assessments that were to be evaluated were examined and for each assessment task that could be rated a rating form was created. An illustrative form is included in Appendix A. Rating forms were developed for all NRTs and all CRTs. For some CRTs assessment tasks could not be rated because they were too vague or because they referred to materials that were not available. For example, for some reading standards district assessments directed teachers to assign students a reading passage and have the students answer specific questions about the passage. Because the passage selection was up to the teacher, it was not possible for the performance level of the task to be determined (different passages would be at different levels of difficulty).

In addition to developing and copying the rating forms, copies of the Power Point presentation that was used for the orientation were made, copies of the performance level descriptions were made, a non-disclosure form, an informed consent form, and a demographic information form were developed and copied, an evaluation form for each meeting was constructed and copied, as were the rating forms for the practice assessments. Copies of all the above materials and forms except for the NRT and CRT rating forms were consolidated into a packet and distributed at the beginning of each meeting. Packets also included a travel reimbursement form.

Copies of the NRTs were obtained from the publishers and all district-provided CRTs were copied and organized for the review process. NRTs and locally developed assessments were not in the packets, but were distributed when it was time to conduct the ratings. This was done for two reasons, the first was security of the materials and the second was that the sheer volume of paper would have been overwhelming for the participants.

Each meeting began with an orientation that described the purpose of the meetings and the context in which the study was being conducted (i.e., an overview of the requirements of NCLB that served as the rationale for conducting the study). Teachers were assured that the performance level descriptions were not being mandated by the state and they were assured that the districts that had volunteered their assessments would be anonymous in the report to the NDE. After the demographic information form, the informed consent form, and the non-disclosure forms were completed, teachers broke into their grade level groups.

The first task of the grade-level groups was to read carefully the performance level descriptions and to discuss them. The objective was for the teachers to understand what these definitions were and how the four performance levels were differentiated in terms of skills, knowledge, and abilities in the respective content areas. This discussion often took up to an hour. This activity was followed by the evaluation and rating of a practice assessment. The Group 1 teachers in phase one who developed the performance level descriptions had rated practice assessment already. These original ratings were not shared with the Group 2 teachers until after their ratings were completed.

The rating process involved examining an assessment task (e.g., a multiple-choice item, a short answer item, or performance type item that is scored using a rubric) and making a decision about the complexity and difficulty of the task. The decision was to be based on the performance level descriptions. Teachers were advised to use the following strategy to make their decision. First, review the performance level description associated with the standard (or in mathematics, the strand) that the assessment task was intended to assess. Then decide if the majority (about 66%) of the beginning students would answer correctly. If not (i.e., the item was too difficult for most Beginning students), then ask if the majority (66%) of the Progressing students would be able to answer correctly. If the assessment task was too difficult for both Beginning and Progressing students, then the process was continued asking about Proficient and if necessary, Advanced students.

For dichotomous items only a single rating was made. If the item was at the Beginning level, it was assumed that students at all higher levels would likely answer correctly. Thus, only the lowest performance level was marked on the rating form. If the assessment

task was a performance task (polytomously scored using a rubric), then teachers determined the appropriate levels of performance for the task given the scoring process. For example, if a task required the student to create a story map, and the rubric provided for up to 5 points for the story map, then teachers evaluated the complexity of the story, the information or structure provided in the task description (e.g., if the elements of the story map were all provided or if the student had to construct the map independently) and the scoring elements as described in the rubric. After evaluating these, the teachers made a judgment about whether Beginning students could get one or more points, then whether Progressing students could get additional points that were unique from the points that the Beginning students achieved. Also indicated was whether there were additional points that might be unique to Proficient or Advanced performance. Thus, for polytomous items, multiple ratings were possible. In some school districts a series of dichotomous items designed to assess a single standard were grouped together by the district so they could be scored using a "rubric." For example a series of 15 computation items might be scored such that students who answered 4 or fewer correct were classified as beginning, whereas those who scored between 5 and 8 correct were progressing, and so on. We chose, in most cases, to abandon the district's scoring method and rate each item independently. The rationale for our decision was that the districts had not looked at each item in terms of the definitions of the performance levels used in this study and we did not want to be bound by the districts determination of performance levels.

The rating process for the practice tests had the teachers discuss and come to consensus on the ratings as a group for the assessment tasks associated with the first standard. For the next standard, teachers read and evaluated each assessment task independently, followed by a discussion that resulted in consensus. This "discussion" entailed the Table Leader (for the practice that person was the Buros facilitator – Impara, Buckendahl, or Jacobson – asking for a "vote" for each performance level for each task. If all panelists agreed on the performance level, there was no discussion. If the majority agreed, then some discussion might be appropriate, depending on the positions taken by those in the minority. Discussion was held to a maximum of 4 minutes for any single assessment task (there were too many tasks to permit unlimited discussion). After consensus was reached, the group facilitator, who served as a table leader, advised the panel what the original rating had been by the Group 1 teachers at the phase one meeting in Lincoln. This process continued until all practice items were rated.

The first operational assessment at each meeting was for the NRT that was evaluated at that location. The evaluation of the NRT was often completed by the end of the first half--day. The CRTs were then evaluated. The Buros facilitator assigned a different person to serve as Table Leader for each assessment. The responsibilities of the Table Leader were to lead the discussion, when discussion was needed, and to complete a Consensus Rating form for the assessment being rated. The purposes for assigning one of the teachers to serve as Table Leader was to eliminate the possibility that the facilitator was "driving" the decisions about each assessment task and to permit the facilitator to organize the materials for the next assessment (the assessment tasks and the rating forms), to look over the rating forms from the previous task, and to be available to respond to questions about the process or the assessment. Having a teacher serve as the Table Leader insured the independence of the process. After all assessment tasks on an assessment were rated, the Consensus Rating form and the individual teacher rating forms were collected and the

next set of materials was distributed and a new Table Leader assigned. This was done for all assessments.

Analyzing the data

This element of the project represents the greatest conundrum. The research literature in the measurement field has not yet dealt with the problem of what constitutes sufficiency of measurement to make a classification decision without benefit of having a criterion variable (i.e., a variable that can be used to determine the concurrent or future level of performance of the examinee). In the late 1960s or early 1970s, Millman indicated that, depending on the criterion level (not to be confused with a criterion variable, in this case a determination that if a student met a criterion of 80% correct he or she was considered "competent") reliable classification decisions could be made with 7 or 8 test questions (assumed to be dichotomous). Millman made no explicit assumptions other than the test questions measured the objective (standard). In a later study (about 1984), Subkoviak indicated that 6 test questions (also dichotomously scored) would be sufficient if the criterion performance level was about 67%. Like Millman, Subkoviak did not specify the nature of the items except that the items were assumed to measure the objective or standard[3]. Moreover, both Millman and Subkoviak were interested in making only a dichotomous decision (master or non master) with some acceptable level of confidence. Thus, depending on some a priori criterion level, that is empirical rather than on some descriptive definition of what skills, knowledge, and ability constitutes mastery (or proficiency or what ever levels of performance are of interest), the number of items varies with the empirical value. Up to now, to the best of our knowledge, no one has undertaken a mathematic derivation to determine the amount of measurement information needed to make multiple classification decision when the classifications are based on constitutive rather than empirical definitions.

In some contexts (e.g., the certification and licensure fields) this lack of data is resolved by having large numbers of assessment tasks and by setting performance standards (cut scores). Thus, if there are large numbers of "easy" tasks, then the standard setters can reflect this by setting a higher standard. Thus, the empirical definition is set after the fact, and it is based on the specific test. The standard setting task might be simplified if assessment tasks were able to be classified in advance as to whether they were likely be answered correctly by individuals at different levels of qualification, but so far this standard setting strategy has not been employed.

Because of the absence of standard setting information for the assessments and because of the lack of measurement theory, a compromise strategy was used in the analysis of the data from phase two of this study. The strategy assumes that each decision is dichotomous. That is, a student is either Beginning or above; a student is either Progressing or below (or above), a student is either Proficient or below (or above), etc. Using this strategy we, have used Subkoviak's suggestion that at least 6 items (or

[3] Both Millman and Subkoviak were operating in the realm of "mastery" learning and were operating under the assumption that objectives/standards were narrowly defined and content focused. This is not the case with many state standards that are broadly defined and not narrowly content focused. The breadth of focus of many standards presents significant measurement problems when it comes to making classification decisions.

measurement opportunities) are needed at each level and that at least 4 points (of 6) must be obtained. Because we are not actually classifying students, the number points obtained is not particularly relevant.

Using this strategy, each data set was analyzed and determinations made about the potential adequacy of the assessments used in this study to classify students into multiple performance level categories. Each NRT was analyzed independently. Each CRT was also analyzed independently. For the CRTs that were evaluated at more than one meeting, two analyses were completed. The first, related to the "sufficiency" of the assessment to classify students, used the average ratings across all meetings to make the determination. The second analysis of these data was done to examine the consistency of ratings across meetings. The second analysis is reported following the other results as though it were a separate study.

Reporting the results

In addition to the phase one report and this report, a modified version of this report will be sent to the publishers that provided their tests (Riverside, Harcourt Educational Measurement, and CTB/McGraw Hill). Publishers will receive a letter indicating which was their test. Similarly, each school district will receive a report with information permitting them to know which was their test.

RESULTS

The results of the study are organized by content area and by grade level. Within these groupings the regional groupings are reported for unique districts within the region. The common districts (those rated in all three regions[4]) are always reported in the East region. Following the tables and discussions of the ratings of assessment tasks are the results of the evaluations of each meeting. These are grouped by subject area and by location. Finally, the results of the common districts are provided by content area, grade level, and region. Comments are also made about the overall consistency of the ratings.

Because Publishers and districts were provided anonymity in this study, they are numbered rather than named. Thus, the NRT that was rated in the eastern region meeting is identified as NRT-1. Similarly, the CRTs provided by districts are similarly identified. For example, a district's assessment rated in the eastern region meeting might be identified as District 1-E. The districts that were evaluated across districts (the common CRTs) are identified as District 1-C.

The tables shown below describe the results of the Grade 4 meetings for Reading.

Reading Grade 4 – East – NRT-1

As shown in Table 3 below, the teachers evaluated NRT-1 as having items in three of the four performance level categories. This NRT has two subtests containing items that were aligned in an earlier study to Standard 4.1.3. A previous study found that, across the two subtests, a total of 32 items were aligned to this standard.

[4] Lodgepole's mathematics assessments were rated in two of the regional meetings. These results are included as though they were rated uniquely at each meeting, that is, it is not treated as a "common" district. The coded designation for Lodgepole is different in each of the two meetings to protect the confidentiality of the district.

Of the 32 test items the preponderance of items are in the Proficient and Progressing performance levels. There are too few items at the Advanced level to feel comfortable in making such a classification for high scoring students. Although there are too few items at the Beginning level to make a classification, by default, if a student answered fewer than 5 – 7 items correctly, they would likely be below Progressing. That is, they would be Beginning. Thus, a student who obtained a score between 7 and 12 would result in classifying a student as being Progressing. Similarly, a student who obtained a score of 16 to 31 would be classified as being Proficient. However, if a student answered all the questions correctly, a classification of Advanced may be reasonable. The teachers were not asked to set "cut scores" that differentiated between performance levels, so the score points that are suggested in this report are based only on the number of items at a particular performance level and the decision rule described above related to the number of items needed to make a classification decision.

Note that there are some grey areas. Students who score higher than 12, but lower than 16 may be either Progressing or Proficient. It would be difficult to make a definitive classification. A decision rule could well be made to classify all such students as being in the higher category. Such a rule would have to be applied consistently across tests, grades, and years.

In summary, this NRT at grade 4 in reading provides (more than) sufficient information to classify students into the Beginning, Progressing, and Proficient performance levels for Standard 4.1.3 in reading. Some students may be classified as Advanced, but such a decision would be tenuous without one or two additional items measuring that performance level.

Table 3. Consensus item ratings from the Eastern regional meeting for NRT-1 in Reading for Grade 4 for Standard 4.1.3.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Reading Comprehension Subtest	31	5	14	12	0
Usage & Expression Subtest	1	0	1	0	0
Total	32	5	15	12	0

As shown in Table 4, there were three unique districts whose assessments were rated in the Eastern regional meeting. These district assessments varied in terms of the number of assessment activities associated with each standard. All of the assessments tasks that were evaluated were dichotomously scored.

The number of assessment opportunities ranged from a low of 5 for Standard 4.1.5 in District 1-E to a high of 50 for Standard 4.1.2 in district 3-E. Some standards had no items rated (shown as NR in the items column). This occurred for one of several reasons: a) the district did not provide the assessment for that standard (e.g., they may be using the NRT to assess Standard 4.1.3), b) they provided an assessment, but it was not constructed in a way that made it amenable to being rated (e.g., the teacher was directed to select an

unspecified reading passage, thus the difficulty of the task would be a function of the passage selected), or c) a reading passage was identified, but not included in the materials provided for this project and the teachers on the panel were not familiar enough with the passage to evaluate the difficulty of the assessment task.

From Table 4, Standard 4.1.1 is being assessed by District 1-E at all levels except Advanced, whereas District 3-E provides assessment at the Progressing level and by default (i.e., students who answer fewer than 5 or 6 items correctly) at the Beginning level. Standard 4.1.2 is also quite variable in terms of the inferences that can be drawn about student performance levels. In District 1-E inferences can be made about Progressing and Beginning (by inference), and if a student answered all items correctly, one might suggest a Proficient classification (but with some caution). District 2-E, however, has too few items to make any classification decision other than possibly a beginning classification if only 1 or 2 items were answered correctly and a Proficient classification if 8 or 10 items were answered correctly. District 3-E, on the other hand has more than sufficient items to make classifications into all categories except Advanced. In District 1-E Standards 4.1.3 and 4.1.4 can provide classification information for all Levels except Advanced and this same level of sufficiency holds for District 3-E for Standard 4.1.4. Only District 3-E has enough measurement information for Standard 4.1.5 to make any classifications and those classification are limited to Advanced, Proficient, or Below Proficient (for students who answer fewer than 5 items correctly). Only District 2-E had assessment tasks that were rated for Standard 4.1.6 and all classification levels can be made.

Overall, these districts provide midrange performance information. Although there are some standards for which a direct or inferential classification of Beginning may be made, there are few for identifying students who are performing at the Advance level.

Table 4. Consensus item ratings from the Eastern regional meetings for CRTs in Reading for Grade 4 for Standards 4.1.1 – 4.1.6 for Eastern districts (non-common)

	District 1-E					**District 2-E**					**District 3-E**				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.1	36	0	8	22	6	NR					40	0	0	40	0
Standard 4.1.2	40	0	4	32	4	10	3	3	3	1	50	0	25	21	4
Standard 4.1.3	17	0	9	6	2	NR					NR				
Standard 4.1.4	26	0	10	9	7	20	0	0	20	0	20	0	15	5	0
Standard 4.1.5	5	0	5	0	0	NR					17	12	5	0	0
Standard 4.1.6	NR					36	6	6	6	6	NR				

Table 5 shows the results across all three regional meetings for the two districts that were in common across the meetings. The analysis is based on the average ratings across the three meetings. It should be noted that District 1-All has assessment tasks that are scored

using a rubric for Standard 4.1.6. In this case the number of ratings will not necessarily add to the number of assessment items because raters provided a rating for each performance level for which an inference could be made for this assessment task. Because averages were taken across the three meetings some fractional values were obtained. These fractions are always rounded up for making an interpretation for this report. The assessment tasks for District 2-All comprised only dichotomously scored items.

In general, both districts provide assessments that can at least differentiate between Proficient and Below Proficient students. One or both districts provide more information for some standards. For example, both districts provide at least three levels of performance classifications for Standard 4.1.2. For District 1-All on Standard 4.1.2, they would likely be able to classify student performance into Proficient, Progressing, and Beginning (by inference) levels. District 2-All for this same standard would likely be able to classify student performance in all four levels, making a similar inference about the Beginning level.

The assessments in these two districts generally provide little information in the two extreme classifications (Advanced and Beginning). Beginning classifications can be made by inference for some standards, but inference will not work for making classifications of Advanced.

Table 5. Consensus item ratings for the district assessments rated in all three regions for all three regions for Grade 4 for Standards 4.1.1 – 4.1.6.

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.1	16	1.7	6.7	6.3	1.3	14	1.7	4.3	4.3	3.7
Standard 4.1.2	16	1.7	6.7	6.3	1.3	22	6.3	9	6.7	0
Standard 4.1.3	16	1.7	6.7	6.3	1.3	18	.7	6.7	9.7	1
Standard 4.1.4	9	0	6	3	0	22	2.7	6.7	8.7	4
Standard 4.1.5	15	0	13.7	1.3	0	11	2	9.3	.7	0
Standard 4.1.6	9	2	3.3	3.3	2	18	8.7	8.7	0	0

Reading Grade 4 – Central

As shown in Table 6 below, the teachers evaluated NRT-2 as having items in all categories of performance levels. This NRT has two subtests containing items that were aligned in an earlier study to Standard 4.1.3. A previous study found that, across the two subtests, a total of 40 items were aligned to this standard.

Of the 40 test items the preponderance of items are in the Proficient and Progressing performance levels, but there are sufficient items at the Advanced level to feel comfortable in making such a classification for high scoring students. Although there are too few items at the Beginning level to make a classification, by default, if a student answered fewer than 5 – 7 items correctly, they would likely be below Progressing. That is, they would be Beginning. Thus, a student who obtained a score between 7 and 11 would result in classifying a student as being Progressing. Similarly, a student who

obtained a score of 15 to 31 would be classified as being Proficient and students scoring 35 or higher would be clearly performing at the Advanced level on this standard.

Note that there are some grey areas. Students who score higher than 11, but lower than 15 may be either Progressing or Proficient. It would be difficult to make a definitive classification. A decision rule could well be made to classify all such students as being in the higher category. Such a rule would have to be applied consistently across tests, grades, and years.

In summary, this NRT at grade 4 in reading provides (more than) sufficient information to classify students into the four performance levels for Standard 4.1.3 in reading.

Table 6. Consensus item ratings from the central regional meeting for NRT-2 in Reading for Grade 4 for Standard 4.1.3.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Reading Comprehension	36	9	13	11	3
Language	4	0	4	0	0
Total	40	9	17	11	3

In the central regional meeting two unique district assessments were evaluated. All of the assessments tasks that were evaluated were dichotomously scored.

As can be seen in Table 7, District 1-C had assessments rated on only three of the six standards. Of these assessments, those associated with standard 4.1.1 had 59 items, and the preponderance of these items were rated at the Beginning and Progressing performance levels. Thus, for this standard, the District 1-C assessment can effectively classify students as Beginning or Progressing. There are too few measurement opportunities to make classifications at the Proficient level (although a student who obtained all 59 points might well be Proficient. District 1-C's assessments of standards 4.1.5 and 4.1.6 can be used to make three levels of classifications: Advanced, Proficient, or below Proficient (by inference if few items were answered correctly).

For District 2-C, five assessments were evaluated. These assessments ranged from 10 to 25 assessment tasks, all of which were dichotomously scored. Of these five assessments, only the assessment for standard 4.1.2 can be used to classify students into all four performance categories. Three others can be used to classify students into three categories. Specifically, the assessments for standards 4.1.3 and 4.1.4 can classify students as Proficient, Progressing, or (by inference) Beginning. The assessment for standard 4.1.5 can be used to classify students as Advanced, Proficient, or Below Proficient. The assessment used for standard 4.1.6 can classify students as either Proficient or Below Proficient.

Most of the assessment tasks that were rated provided only middle level classification information. Beginning level classifications were possible (with one exception) only by inference. There are also few opportunities to be comfortable in classifying students as Advanced from the data provided in these assessments.

Table 7. Consensus item ratings from the central regional meetings for CRTs in Reading for Grade 4 for Standards 4.1.1 – 4.1.6 for Eastern districts (non-common)

	District 1-C					District 2-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.1	59	0	3	23	33	NR				
Standard 4.1.2	NR					22	6	7	8	1
Standard 4.1.3	NR					10	0	4	6	0
Standard 4.1.4	NR					25	0	10	15	0
Standard 4.1.5	12	5	6	1	0	15	6	8	1	0
Standard 4.1.6	17	4	11	1	1	14	0	10	0	4

In Table 8 are the results from the Western regional meeting for the NRT that was examined at that site. This NRT had one subtest containing 42 items related to standard 4.1.3. This test can be used to classify students into all four performance categories, but the number of items classified as Advanced is marginal for this classification. The determination of a student as Beginning is by inference, rather than because the number of items at the Beginning level is sufficient.

Table 8. Consensus item ratings from the Western regional meeting for NRT-3 in Reading for Grade 4 for Standard 4.1.3.

Subtest	**Items.**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Reading Language Arts	42	5	23	10	4

Table 9 shows the results of the three assessment evaluations that were unique to that region. All of the assessment tasks that were evaluated were dichotomously scored.

In general, District 1-W's assessments can be used to classify students as either Proficient or Below Proficient. The exception is the assessment for standard 4.1.5, which provides sufficient measurement information to classify students as either Advanced, Proficient, or Below Proficient.

Of the five assessments for District 2-W, two can be used to classify students into three categories. The assessment standards4.1.1 can be used to classify students into Proficient, Progressing, or Beginning (by inference) and the assessment for standard 4.1.4 can be used to classify students into Advance, Proficient, or Below Proficient categories. The only other assessment that can be used for classification is for standard 4.1.5, which can classify students as either Advanced or Below Advanced. The remaining two assessments do not have enough items to provide any accurate classification information.

District 3-W's assessments that were evaluated all provide sufficient information to make two classification decisions. Specifically the assessments for all standards can be used to

infer that students are Proficient or Below Proficient. These classifications are, for some standards highly inferential. For example, for standard 4.1.1, if a student answered as many as five items correctly, he or she would have answered correctly at least three items at the Proficient level and one item at the Advanced level, suggesting the student is at least proficient. Students answering fewer than five items correctly would be classified as Below Proficient.

Note that there were virtually no items provided at the Beginning level in these districts for any standard. All classifications of Beginning would have to be made by inference when there were enough assessment tasks provided at the Progressing level. Similarly, there are only a limited number of standards for which these districts have provided opportunities for Advanced students to demonstrate their higher-level skills.

Table 9. Consensus item ratings from the Western regional meetings for CRTs in Reading for Grade 4 for Standards 4.1.1 – 4.1.6 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.1	10	3	6	1	0	21	3	9	9	0	8	4	3	1	0
Standard 4.1.2	10	3	6	1	0	NR					8	4	3	1	0
Standard 4.1.3	10	3	6	1	0	4	0	4	0	0	10	4	6	0	0
Standard 4.1.4	10	0	8	2	0	30	9	19	2	0	NR				
Standard 4.1.5	15	9	5	1	0	8	8	0	0	0	10	0	10	0	0
Standard 4.1.6	10	1	7	2	0	7	2	4	1	0	10	0	7	3	0

Reading Grade 8

The NRT that was evaluated in the Eastern regional meeting included items related to standard 8.1.1 in four separate subtests. Only two of these subtests have more than one item, but the four subtests combined have a total of 82 items aligned to this standard. The teachers who rated these items found none at the Beginning level. However, there are sufficient items at the higher performance levels so that classification can be made at all four levels (the beginning level by inference for students who answer fewer than 5 or 6 items correctly).

Table 10. Consensus item ratings from the Eastern regional meeting for NRT-1 in Reading for Grade 8 for Standard 8.1.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Reading Comprehension	49	9	27	13	0
Usage & Expression	1	1	0	0	0
Maps & Diagrams	31	2	17	12	0
Reference Materials	1	0	1	0	0
Total	82	12	45	25	0

In Table 11 the results for the three unique districts are shown. As can be seen in this table, all three districts had a wide range of assessment tasks related to each of the six standards. The lowest number of assessment tasks is found in District 2-E where there are only four tasks associated with standard 8.1.2. This same district had 66 assessment tasks associated with standard 8.1.5. All assessments that were rated in these three districts were dichotomously scored tasks (no rubrics were used for scoring).

In District 1-E performance level inferences can be made only for standards 8.1.1 through 8.1.4. Standards 8.1.5, and 8.1.6 do not have enough items at any performance level to make reasonable inferences about student performance. For the standards about which performance level inferences are reasonable, students can be classified as Proficient or below for standard 8.1.1, as Progressing or below for standard 8.1.2. For standards 8.1.3 and 8.1.4, students may be classified as Proficient, Progressing, or Beginning (by inference), although the classification of Proficient for standard 8.1.3 is questionable (a student would have to answer all items correctly to make that classification).

District 2-E has enough assessment tasks to make two or more proficiency classifications for all standards except 8.1.2. Students can be classified as Proficient, Progressing, or Beginning for standards 8.1.1, 8.1.3, 8.1.4, and 8.1.5. For standard 8.1.6 students can be classified as Progressing or Beginning. Note that no assessment tasks were classified as Advanced for any of these standards for District 2-E.

District 3-E had no assessments rated for standards 8.1.5 and 8.1.6. The assessments for the remaining standards permit classifying students into three performance levels for standards 8.1.1, 8.1.2, and 8.1.3 and into all four performance levels for standard 8.1.4. The performance levels associated with the first three standards are Proficient, Progressing, and Beginning (by inference).

In general, these districts tend to have assessments that may be used for classifying students into all levels except Advanced. There were very few items rated as Advanced for any of the assessments for these three districts. For the classification of Beginning, most decisions will be made by inference.

Table 11. Consensus item ratings from the Eastern regional meetings for CRTs in Reading for Grade 8 for Standards 8.1.1 – 8.1.6 for Eastern districts (non-common)

	District 1-E					District 2-E					District 3-E				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.1	13	0	12	1	0	60	0	14	22	24	13	1	5	7	0
Standard 8.1.2	12	0	3	7	2	4	0	0	4	0	16	0	8	7	1
Standard 8.1.3	15	0	5	8	2	59	0	9	48	2	18	0	10	8	0
Standard 8.1.4	27	0	22	5	0	35	0	8	21	6	29	9	10	10	0
Standard 8.1.5	5	0	2	2	1	66	0	12	15	39	NR				
Standard 8.1.6	8	0	4	4	0	16	0	0	8	8	NR				

The ratings for the two common districts are shown in Table 12. As was the case for the 4th grade assessments that were rated in all three of the regional meetings, the ratings for these common districts were averaged across the three meetings to obtain an overall consensus. Thus, the number of assessment tasks for each standard are may be fractional values and they may not add to the total number of items due to rounding. Moreover, in District 1-All for standards 8.1.1, 8.1.2, 8.1.3, 8.1.4, and 8.1.5 some assessment tasks are rubric scored. The five items used to assess standard 8.1.6 are not rubric scored. Similarly, in District 2-All for standards 8.1.3, 8.1.4 and 8.1.5, some assessment tasks were not rated in one meeting (teachers either felt that the tasks were not aligned to the standard, thus ratings would be misleading, or teachers were not familiar enough with the materials to make rating decisions). For the standards where rubrics are used to score items the sum of the values across the performance levels will exceed the number of items for those standards (because raters may have indicated that more than one classification decision could be made from the task and the scoring). For the standards that were not rated in all locations, the number of assessment tasks may be fewer than the total, because they were rated in only two of the three meetings (this tends to underestimate the extent that performance level classifications can be made).

In District 1-All, students can be classified into all four performance categories only on standard 8.1.4. Students can be classified as being Proficient, Progressing, or Beginning for standards 8.1.1, 8.1.2, and 8.1.3. For standard 8.1.5 students can be classified only as either Progressing or Beginning. There is not enough measurement information available to make a classification decision other than Beginning for standard 8.1.6, and such a classification should be done cautiously.

The extent to which performance level classifications can be made across the five standards that were rated for District 2-All is somewhat limited. There are no standards for which a classification of Advanced can be made. Classifications of Proficient can be made for four standards (8.1.1, 8.1.3, 8.1.4, and 8.1.5). Classifications of Progressing can be made for standards 8.1.1, 8.1.2, and 8.1.5. The Beginning classification can be made for standards 8.1.1, 8.1.2, and 8.1.5 (by inference).

As has been the case for other assessments, there tends to be a paucity of assessment opportunities at the Advanced level of performance. In general the ability of these assessment to assign performance level classifications is strongest in the two middle

categories, with classifications of Beginning being made sometimes directly, but mostly by inference for students who obtain zero or very low scores.

Table 12. Consensus item ratings for the district assessments rated in all three regions for Grade 8 for Standards 8.1.1 – 8.1.6.

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.1	22	1.67	12	11.33	8.33	27	2.67	5.33	9	10
Standard 8.1.2	15	0	4.67	11.33	2.67	21	1.33	1.33	7.33	11
Standard 8.1.3	23	10	14.33	6.67	6.33	28	2.67	16.33	2.67	1.67
Standard 8.1.4	20	1.33	4.67	10.33	5.67	18	.33	10.67	3.33	.33
Standard 8.1.5	10	0	1	6.67	5.33	19	0	6.67	7	1.33
Standard 8.1.6	5	0	0	0	5	NR				

The NRT that was evaluated in the central regional meeting included items related to standard 8.1.1 in two distinct subtests. One of these subtests had only three items that aligned to the appropriate standard (8.1.1), but the two subtests combined have a total of 47 items aligned to this standard. The teachers who rated these items found sufficient items at all performance levels so that classification can be made at each of the four levels.

Table 13. Consensus item ratings from the central regional meeting for NRT-2 in Reading for Grade 8 for Standard 8.1.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Reading Comprehension	44	13	16	6	9
Language	3	1	1	1	0
Total	47	14	17	7	9

In the three unique districts whose assessments were rated in the Western regional meeting it should be noted that each district had at least one standard for which no assessment was rated. For the standards that were rated, there was a moderately wide range of assessment tasks across standards. The range was from a low of six to a high of 27. All assessment tasks in these three districts were dichotomously scores (no rubric scoring was employed).

In District 1-C, no assessments were rated for standards 8.1.1 or 8.1.2. Of the remaining four standards, there was sufficient measurement information to classify students as Advance, Proficient, or Below Proficient for standard 8.1.3. For standard 8.1.4, classifications of Proficient, Progressing, or Beginning (by inference) were possible. The assessment for standard 8.1.5 permits classifications of Progressing or Beginning (by inference), whereas there is not enough information available from the assessment of standard 8.1.6 to make any classifications.

District 2-C also has two standards that had assessments that were not rated, standards 8.1.5 and 8.1.6. Standards 8.1.1 and 8.1.2, each has sufficient assessment information to classify students into two categories. For 8.1.1, students can be put into either Proficient or Below Proficient categories and for 8.1.2 students may be classified as either Progressing or Beginning (by inference). The assessments for 8.1.3 and 8.1.4 each permit classification into Proficient, Progressing, or Beginning (by inference), although caution is suggested for making the Progressing classification for standard 8.1.3 (a score of about six may justify such a classification).

The assessment for only one standard (8.1.6) was not rated for District 3-C. Of the five standards for which assessments were rated, only two provided enough information to assign students to any performance classifications. These two are for standards 8.1.3 and 8.1.4. The assessment for standard 8.1.4 permits classification of students as either Proficient or Below Proficient, whereas the assessment for standard 8.1.4 permits classifications of Proficient, Progressing, and Beginning (by inference).

As has been the case for district assessments described above, there is a general lack of assessment opportunities at the Advanced level. Similarly, there are relatively few assessment tasks at the Beginning level, because assignment to the Progressing level is possible in some cases the Beginning level can be inferred for students who obtain few of the available points on these assessments.

Table 14. Consensus item ratings from the Central regional meetings for CRTs in Reading for Grade 8 for Standards 8.1.1 – 8.1.6 for Central districts (non-common)

	District 1-C					District 2-C					District 3-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.1	NR					16	2	8	3	3	12	4	4	1	3
Standard 8.1.2	NR					6	0	0	6	0	11	4	0	7	0
Standard 8.1.3	12	6	6	0	0	11	0	6	5	0	16	0	16	0	0
Standard 8.1.4	16	0	7	9	0	20	0	6	14	0	27	0	12	15	0
Standard 8.1.5	9	0	1	8	0	NR					6	0	2	3	1
Standard 8.1.6	8	0	0	4	4	NR					NR				

Unlike the previously rated NRTs, this one does not provide a wide distribution of items across performance levels. Note that there are two subtests, one with 35 items aligned to standard 8.1.1 and the second with five aligned items. Of the 43 total items, the preponderance of items are at the Advanced performance level, and 14 are at the Proficient level, leaving the remaining 8 items distributed evenly across the Progressing and Beginning levels. Thus, student classification into three categories is reasonable to make: Advanced, Proficient, and Below Proficient.

Table 15. Consensus item ratings from the Western regional meeting for NRT-3 in Reading for Grade 8 for Standard 8.1.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Comprehension	35	16	12	3	4
Language Arts	8	5	2	1	0
Total	43	21	14	4	4

In the Western regional meeting, one district had only two of its assessment rated. In this case, for District 2-W for the two of its assessments rated, each provided assessment information sufficient for rating students as Beginning on the two relevant standards (8.1.3 and 8.1.4).

District 1-W had between 9 and 19 assessment tasks (all dichotomously scored) for each standard. For standard 8.1.1 the 19 assessment tasks were distributed across all performance levels, thus providing only limited opportunities to confidently classify students. However, students could be confidently classified as Proficient or Below proficient, with the possibility of making some cautious decisions about Progressing and Beginning students. In examining the assessments for standards 8.1.2, 8.1.3, 8.1.4, 8.1.5 and 8.1.6 one could be comfortable in classifying students as Proficient or Below Proficient. (For standard 8.1.3 a student who answered 9 or 10 items correctly might well be at least Proficient. Similarly for the assessment associated with standard 8.1.6, a student obtaining 8 or 9 of the 9 possible points might well be classified as Proficient.).

One standard did not have an assessment rated in District 3-W (standard 8.1.2). Of the remaining standards only three assessments permitted classifying students. The assessments for standards 8.1.1 and 8.1.5 each had six items, but they were distributed across the performance levels in such a way as to preclude making any student performance level classifications. The assessment for standard 8.1.6 had all of its items classified as being at the Beginning level, thus permitting only that classification. The assessments for standards are difficult to judge in terms of classifications. It is clear that they are suitable for classifying students as Advanced, Progressing, or Beginning (by inference). However, it is not clear if these assessments could be used to make a classification of Proficient (e.g., a student who answers all but six or seven items correctly could be either a very high level of Progressing, Proficient, or a low Advanced).

Table 16. Consensus item ratings from the Western regional meetings for CRTs in Reading for Grade 8 for Standards 8.1.1 – 8.1.6 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.1	19	3	6	5	5	NR					6	0	2	3	1
Standard 8.1.2	9	0	6	3	0	NR					NR				
Standard 8.1.3	14	5	4	4	1	12	0	0	0	12	24	12	0	12	0
Standard 8.1.4	13	0	11	1	1	6	0	0	0	6	22	11	0	11	0
Standard 8.1.5	14	0	12	2	0	NR					6	0	2	3	1
Standard 8.1.6	9	1	5	3	0	NR					10	0	0	0	10

Reading High School

The NRT that was evaluated for standard 12.1.1 in the Eastern regional meeting included items in only one subtest. There are 33 items in this subtest related to this standard. The teachers who rated these items found none at the Beginning level. However, there are sufficient items at the higher performance levels so that classification can be made at all four levels (the beginning level by inference for students who answer fewer than 5 or 6 items correctly).

Table 17. Consensus item ratings from the Eastern regional meeting for NRT-1 in Reading for High School for Standard 12.1.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Reading Comprehension	33	6	19	8	0
Total	33	6	19	8	0

For District 1-E, there are too few items at any single performance level to make confident classifications for standards 12.1.1 or 12.1.2. However, for standard 12.1.2 an inference that a student is Below Proficient might be possible for students answering only 4 or fewer items correctly and a student who answers 6 or more items correctly might well be classified as being Proficient. The assessments for the remaining standards permit classification into two levels. These two levels for standards 12.1.3, 12.1.4, and 12.1.5 are Proficient and Below Proficient (by inference) and for standard 12.1.6 the two levels are Progressing and Beginning (by inference).

District 2-E had only three assessments rated. These assessments were for standards 12.1.2, 12.1.3, and 12.1.4. The assessments for standards 12.1.2 and 12.1.4 were rubric scored, thus the assessments may be sufficient for classifying students at all levels. This is because the teachers found the assessment and the rubric permitted classification at all four performance levels. The assessment for standard 12.1.3 had 10 dichotomously scored items and all were judged to be at the Progressing levels, permitting classification at only the Progressing and Beginning (by inference) levels.

In District 3-E, only three assessments were evaluated. Each of these assessments consisted of 15 dichotomously scored items that are distributed evenly across the three lowest performance levels. Thus, it is difficult to make confident decisions about student performance levels except by inference. For example, a student who answered 8 – 10 items correctly might be classified at the Progressing level and one who answered fewer than 6 items correctly might be classified as Beginning. A student who answered all 15 items correctly may well be Proficient, but could also be at a high level of Progressing. There are essentially no opportunities for an Advanced student to demonstrate that level of performance.

Table 18. Consensus item ratings from the Eastern regional meetings for CRTs in Reading for High School for Standards 12.1.1 – 12.1.6 for Eastern districts (non-common)

	District 1-E					District 2-E					District 3-E				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.1	11	0	3	5	3	NR					NR				
Standard 12.1.2	10	2	5	3	0	12	3	3	3	3	NR				
Standard 12.1.3	10	0	6	4	0	10	0	0	10	0	NR				
Standard 12.1.4	10	0	9	1	0	4	1	1	1	1	15	0	5	5	5
Standard 12.1.5	10	0	7	3	0	NR					15	0	5	5	5
Standard 12.1.6	14	0	2	12	0	NR					15	0	5	5	2

There were only two common districts that provided assessments for high school. These were Districts 1-All and 2-All.

District 1-All had from 18 to 43 dichotomously scored items for the high school reading standards. However, these items were not distributed across all performance levels for all standards, thus for most standards at most three performance level classifications are possible. Specifically, for standard 12.1.1 students could readily be classified as Progressing or Beginning. Students who answered all 22 items correctly might be classified as Above Progressing (but they could also be high Progressing). The assessment for standard 12.1.2 permits classifying students into only two levels of performance, Progressing or Beginning. Students can be classified confidently into three performance levels for standard 12.1.3 and with some trepidation into the highest performance level (if they answered virtually all of the 50 items correctly). The remaining assessments for standards 12.1.4 through 12.1.6 have sufficient assessment opportunities to classify students as Proficient, Progressing, or Beginning.

Table 19. Consensus item ratings for the district assessments rated in all three regions for Standards 12.1.1 – 12.1.6 (common districts)

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.1	22	1	4	10	7	11	0	0	3.7	7.3
Standard 12.1.2	18	0	1.3	6.7	10	9	0.7	2	3.7	2.7
Standard 12.1.3	50	5	36.3	5.7	3	16	0	4.7	9	2.3
Standard 12.1.4	22	0	6.7	9	6.3	9	1.3	4.3	2.3	1
Standard 12.1.5	24	0	6	16.7	1.3	19	0.3	4.7	12.3	1.3
Standard 12.1.6	43	0.7	7	22	13.3	23	0	0	9.7	13.3

The NRT evaluated in the Central regional meeting had 42 multiple-choice items in a single subtest that were aligned with standard 12.1.1. These 42 items were predominantly in the three lowest performance levels. There are, however, five items that were designated by the teachers as being at the Advanced level, so students who answer all 42 items correctly, might be either at the Advanced or highly Proficient performance level.

Table 20. Consensus item ratings from the Central regional meeting for NRT-2 in Reading for High School for Standard 12.1.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Reading Comprehension	42	5	10	18	9
Total	42	5	10	18	9

Two districts' assessments were rated at the Central regional meeting for the six reading standards. In District 1-C only four assessments were rated and of these four, the assessments for three standards (standards 12.1.3, 12.1.4, and 12.1.5) were found to be appropriate for classifying students into two performance categories, while the fourth assessment (standard 12.1.6) has items that are sufficient for classifying students into the Beginning performance level. The assessment associated with standard 12.1.3 has sufficient items at the Progressing level to make classifications at that level and by inference at the Beginning level. The assessments for standard 12.1.4 has sufficient measurement breadth to make classifications at the Progressing or Below Progressing

level, whereas the assessment for standard 12.1.5 can be used to classify students as either Proficient or Beginning. It may be reasonable to consider students who answered 15 – 17 items correctly as being Progressing, but such a classification should be made with caution.

Table 21. Consensus item ratings from the Central regional meetings for CRTs in Reading for High School for Standards 12.1.1 – 12.1.6 for Central districts (non-common)

	District 1-C					District 2-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.1	NR					7	0	6	0	1
Standard 12.1.2	NR					15	0	1	0	14
Standard 12.1.3	17	0	0	17	0	16	0	6	10	0
Standard 12.1.4	13	2	6	4	1	25	0	0	3	22
Standard 12.1.5	24	1	7	3	13	25	0	0	3	22
Standard 12.1.6	19	0	0	5	14	10	0	0	0	10

The NRT that was evaluated in the Western regional meeting has 42 multiple-choice items that were aligned with standard 12.1.1. These 42 items are distributed across all performance levels and provide sufficient measurement information to classify students in all four performance levels.

Table 22. Consensus item ratings from the Western regional meeting for NRT-3 in Reading for High School for Standard 12.1.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Reading Comprehension	42	9	12	14	7
Total	42	9	12	14	7

District 1-W had assessments rated for all six standards. Only the assessment for standard 12.1.2 did not have sufficient breadth of item difficulty to make any performance level classifications. The remaining assessments provided enough items to make classifications of either Progressing or Beginning (by inference).

In District 2-W, assessments for two standards were not rated (standards 12.1.2 and 12.1.6). The assessment for standards 12.1.1 and 12.1.5 can be used to classify students at the Beginning level. The assessment for standard12.1.3 had 15 items at the Proficient level and thus permits classification of students as Proficient or Below Proficient. Although the assessment for standard 12.1.4 had 11 total items, these items are too widely distributed for making any classification decisions (except that a student who answered all items correctly may be cautiously classified as being Proficient and a student who answered only 2 or 3 items correctly may be at the Beginning level).

The assessments for District 3-W had some items for all standards, but the only standards for which classification decisions are possible are standards 12.1.4 (Beginning), 12.1.5 (Progressing or Beginning – by inference), and 12.1.6 (Beginning).

Table 23. Consensus item ratings from the Western regional meetings for CRTs in Reading for High School for Standards 12.1.1 – 12.1.6 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.1	14	0	2	7	5	11	0	0	2	9	6	0	0	4	2
Standard 12.1.2	7	0	2	2	3	NR					5	1	0	1	3
Standard 12.1.3	7	0	0	6	1	15	0	15	0	0	5	0	2	1	2
Standard 12.1.4	14	1	3	9	1	11	0	5	3	3	10	0	2	3	5
Standard 12.1.5	14	1	3	9	1	8	0	1	2	5	10	0	2	7	1
Standard 12.1.6	14	0	2	10	2	NR					10	0	0	3	7

Mathematics Grade 4

Table 24 provides results of the teachers' ratings of NRT-1 that was evaluated in the Eastern regional meeting. This NRT had multiple-choice items that assessed standards 4.2.1 and 4.5.1. As shown in the table, there are 53 items across three subtests that aligned with standard 4.2.1. These items are principally associated with the Proficient and Progressing performance levels. Only 5 items were classified as being at the Advanced level, thus making a performance classification of Advanced questionable. A classification of Beginning could be made by inference for students who answer fewer than five or six items correctly. For standard 4.5.1 there are 10 items in the Mathematics Computation subtest. Of these 10 items, eight were classified at the Proficient level, suggesting that students could be classified as either Proficient or Below Proficient (by inference).

Table 24. Consensus item ratings from the Eastern regional meeting for NRT-1 in Mathematics for Grade 4 for Standards 4.2.1 and 4.5.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Math Concepts and Estimation					
Standard 4.2.1	13	1	11	1	0
Math Problem Solving and Data Interpretation					
Standard 4.2.1	13	2	8	3	0
Standard 4.5.1	10	2	8	0	0
Math Computation					
Standard 4.2.1	27	2	12	13	0
Overall					
Standard 4.2.1	53	5	31	17	0
Standard 4.5.1	10	2	8	0	0

The three districts' CRTs that were rated in the Eastern regional meeting included one district (Lodgpole) that was also rated in the Western regional meeting. This district is not included in the common districts, because that category is reserved for districts rated in all three of the regional meetings. As shown in Table 25, District 1-E had between six

and 23 items associated with the relevant standards. Three of the assessments (for standards 4.1.5, 4.3.4, and 4.6.2) were not broad enough to make any performance level classifications. For the remaining three standards, the assessment for standard 4.2.1 provided sufficient evidence to classify students as Advance, Proficient, or Below Proficient (by inference); the assessments for standards 4.4.2 and 4.5.1 each permit classifications of Proficient and Below Proficient (and for standard 4.5.1 Beginning).

Table 25. Consensus item ratings from the Eastern regional meetings for CRTs in Mathematics for Grade 4 for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Eastern districts (non-common)

	District 1-E					District 2-E					District 3-E				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.5	6	0	2	4	0	6	1	1	3	1	16	1	4	9	2
Standard 4.2.1	21	6	11	4	0	26	1	17	8	0	25	2	22	1	0
Standard 4.3.4	6	0	3	3	0	1	0	1	0	0	30	10	11	9	0
Standard 4.4.2	10	0	5	3	2	19	0	5	10	4	24	0	14	10	0
Standard 4.5.1	23	1	10	8	4	20	1	13	4	3	25	2	18	5	0
Standard 4.6.2	8	3	4	1	0	8	0	4	4	0	25	7	10	6	2

The two districts that had their assessments rated at all three regional meetings each had one standard for which no assessments were included. Specifically, in District 1-All no assessments were rated for standard 4.1.5, whereas for District 2-All no assessment was rated for standard 4.6.2. The other five standards in both districts had wide ranges of measurement opportunities from as few as four items (District 1-All and District 2-All for standard 4.3.4) to 47 items for standard 4.2.1 in District 1-All.

In District 1-all only two standards provided enough measurement information to make differential student classifications. Specifically, for standards 4.2.1 and 4.6.2 the 47 items and 17 items, respectively, permit classifying students as Advanced, Proficient, or Below Proficient. The assessments for the remaining standards do not provide enough diversity of difficulty or enough items to permit differential classification across performance categories.

The situation for District 2-All is slightly more promising than for District 1-All, in that there are three standards for which classifications can be made. The assessment for standard 4.2.1 provides sufficient information for classifying students as Advanced, Proficient, or Below Proficient. For standard 4.4.2 a classification of either Proficient or Below Proficient is possible, whereas for standard 4.5.1 a classification of either Progressing or Beginning can be made based on the assessments that were rated.

Table 26. Consensus item ratings for the district assessments rated in all three regions for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 (common districts).

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.5	NR					10	2.7	4	1.7	1.7
Standard 4.2.1	47	15	27.7	4.3	0	28	18.7	7.3	2	0
Standard 4.3.4	4	0	2	2	0	4	0.7	2.3	1	0
Standard 4.4.2	10	3.7	1.7	2.7	2	17	2.7	9	4	1.3
Standard 4.5.1	7	1	2.3	3.3	0.3	28	4.3	15.3	6.3	2
Standard 4.6.2	17	5.3	10.3	0.7	0.7	NR				

As shown in Table 27, there are two subtests in which items were aligned with standards 4.2.1 and 4.5.1. For standard 4.2.1 there are 33 items and for standard 4.5.1 there are 12 items. The 33 items associated with standard 4.2.1 are concentrated in the Progressing and Proficient categories, thus permitting classification of students as being Proficient, Progressing, or Beginning (by inference). The items aligned with standard 4.5.1 also permit three levels of classification, but these levels are Advance, Proficient, and Below Proficient (by inference).

Table 27. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 4 for Standards 4.2.1 and 4.5.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Math Concepts and Problem Solving					
Standard 4.2.1	11	2	8	1	0
Standard 4.5.1	12	6	6	0	0
Math Computation					
Standard 4.2.1	22	0	12	10	0
Overall					
Standard 4.2.1	33	2	20	11	0
Standard 4.5.1	12	6	6	0	0

In the Central region, with one exception, all standards had assessments that were rated. The exception was District 3-C for standard 4.2.1 for which no assessment was evaluated. The number and scope of assessment tasks that were rated varied widely within an across the districts. In District 1-C, standard 4.1.5 had only a single assessment task that did not provide enough information to make any performance classifications. In that district there was also a standard, 4.2.1 for which there are 120 assessment tasks (distributed across all performance categories) permitting classification into the four performance levels. The remaining assessments permit classification into two performance categories: standard 4.3.4 permits classification of students as either Progressing or Beginning (by inference); standard 4.4.2 permits classification as either Advanced or Below Advanced; and the assessments for standards 4.5.1 and 4.6.2 permit classification of students into Proficient or Below Proficient categories.

The assessments for District 2-C, provide some classification information for all standards (but note that any inferences made for standard 4.4.2 should be with extreme caution). The assessment for standard 4.1.5 permits classification of students as either Advance or Not Advanced. Standard 4.2.1's assessment, with 40 items, permits classification of students into all categories except Advanced, as does the assessment for standard 4.3.4. In contrast to the latter two assessments, the assessments for standards 4.4.2 and 4.5.1 permit classification of students as either Proficient or Below Proficient. Students may be classified as either Advanced, Proficient, or Below Proficient by the assessment used for standard 4.6.2.

Table 28. Consensus item ratings from the Central regional meetings for CRTs in Mathematics for Grade 4 for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Central districts (non-common)

	District 1-C					District 2-C					District 3-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.5	1	0	1	0	0	11	6	1	2	2	14	12	1	1	0
Standard 4.2.1	120	12	34	30	44	40	3	21	10	6	NR				
Standard 4.3.4	10	0	2	8	0	16	0	6	10	0	12	0	7	5	0
Standard 4.4.2	12	6	3	3	2	7	2	5	0	0	16	7	4	4	1
Standard 4.5.1	10	2	7	1	0	15	1	10	4	0	20	10	10	0	0
Standard 4.6.2	10	0	10	0	0	19	7	8	3	1	8	3	5	0	0

NRT-3, which was the initial focus of attention in the Western regional meeting had a total of 35 items that had been aligned with standard 4.2.1 and 12 items that were aligned with standard 4.5.1. The 35 items that measure standard 4.2.1 are distributed across the three highest performance levels and there are enough items at each level to make performance level classifications into all four categories (placement into the Beginning level is by inference for students who fewer than 11 items correctly. For standard 4.5.1 there are only 8 items and they do not provide sufficient information to classify students into all four levels. They do, however, permit students to be classified as Proficient (students who answer most of the items correctly) or Below Proficient (for students who answer fewer than five or six items correctly).

Table 29. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 4 for Standards 4.2.1 and 4.5.1.

Subtest	**Items**	**Advanced**	**Proficient**	**Progressing**	**Beginning**
Mathematics Subset					
Standard 4.2.1	21	5	9	7	0
Standard 4.5.1	12	3	5	0	0
Mathematics Computation					
Standard 4.2.1	14	4	6	4	0
Overall					
Standard 4.2.1	35	9	15	11	0
Standard 4.5.1	8	3	5	0	0

In the Western region, District 1-W's assessments provide some opportunities to classify students into tow or more performance levels for all six standards. The assessment for standard 4.1.5 permits classification of students as either Progressing or Beginning (by inference). The other assessment that permits only two levels of performance classification is for standard 4.5.1, which permits assigning students to either the Proficient or Below Proficient categories. The assessments for standards 4.2.1 and 4.3.4 permits classifying students into three performance levels, Advanced, Proficient, or Below Proficient and Proficient, Progressing, or Beginning (by inference), respectively. The assessments for both standards 4.3.4 and 4.6.2 permit classification into all four performance levels, but the Proficient classification for standard 4.4.2 should be made with extreme caution, because it is requires a high degree of inference.

Five of the assessments for District 2-W permit comfortable classification of students into two or more performance levels. These are the assessments for standards 4.2.1 (classifies students as Proficient or Below Proficient), 4.3.4 (classifies students as Progressing or Beginning – by inference), 4.4.2 classifies students as Proficient or Below Proficient), 4.5.1 (classifies students as Progressing or Beginning – by inference), and 4.6.2 (classifies students as Proficient or Below).

District 3-W's assessments of these standards provide only two opportunities to classify students into two or more performance categories. Specifically, the assessment for standard 4.2.1 permits classifying students as either Proficient or Below Proficient as does the assessment for standard 4.4.2. The remaining assessments do not have sufficient measurement information to make any performance level classifications with any degree of confidence. (Note that an argument could be made for the assessment of standard 4.6.2 that it permits classifying students as Progressing or Below Progressing for students who answer all eight items correctly.)

Table 30. Consensus item ratings from the Western regional meetings for CRTs in Mathematics for Grade 4 for Standards 4.1.5, 4.2.1, 4.3.4, 4.4.2, 4.5.1, and 4.6.2 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 4.1.5	16	1	4	10	1	5	0	2	3	0	5	0	1	4	0
Standard 4.2.1	25	10	13	2	0	10	0	7	3	0	18	4	11	3	0
Standard 4.3.4	30	2	11	17	0	10	0	3	7	0	3	0	0	3	0
Standard 4.4.2	24	11	4	8	1	15	2	9	4	0	5	2	1	1	1
Standard 4.5.1	25	1	21	3	0	16	0	3	13	0	6	0	6	0	0
Standard 4.6.2	25	6	6	12	1	11	3	5	3	3	8	0	4	4	0

Mathematics Grade 8

As shown in Table 31, NRT-1 has one subtest that includes only 14 items related to each of the two standards (8.2.2 and 8.5.2) appropriate to this study. The preponderance of items associated with standard 8.2.2 are at the Proficient level of performance, suggesting that students can be classified as either Proficient or Below Proficient on this standard. For standard 8.5.2, however, the items are distributed across two performance levels, thus providing an opportunity to classify students as Proficient, Progressing, or Beginning (by inference).

Table 31. Consensus item ratings from the Eastern regional meeting for NRT-1 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Math Problem Solving and Data Interpretation					
Standard 8.2.2	14	1	10	3	0
Standard 8.5.2	14	1	6	6	1
Overall					
Standard 8.2.2	14	1	10	3	0
Standard 8.5.2	14	1	6	6	1

The three districts assessments rated in the Eastern region included one district that was also rated in the Western region, as was the case for the grade 4 assessments. Only one standard in one district was not rated. For most of the standards, the number of items was greater than 20. Most of the exceptions to this general statement are seen in District 1-E, which had more than 20 items only for standards 8.4.1 and 8.5.2.

In District 1-E the 12 assessment tasks that focus on standard 8.1.4 are divided evenly across the Proficient and Progressing categories, thus permitting classifying students into these two categories and into the Beginning category (by inference). Measurement opportunities are more limited for standards 8.2.2 and 8.3.2 such that only classifications of Progressing or Beginning (by inference) can be made. The large number of assessment tasks associated with standards 8.4.1 and 8.5.2 are distributed across all performance classifications, permitting students to be classified into all four performance levels. The five items for standard 8.6.3 are not sufficient to make any classification decision.

The one standard for which there were no items rated was standard 8.2.2 for District 2-E. In District 2-E, although most standards had between 10 and 40 items, the assessments that were rated provided no information at the Advanced level. The assessments for standards 8.1.4, 8.3.2, 8.4.1, and 8.5.2 provided sufficient information to place students into the Proficient, Progressing, and Beginning (by inference) levels. The remaining assessment, for standard 8.6.3, permits assigning students to Proficient or Below Proficient levels.

Assessments for all six standards were rated for District 3-E and these assessments all permitted multiple classifications. For standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, and 8.5.2 the assessments permit classification into Proficient, Progressing, and Beginning levels. As is often the case the Beginning level classification is made by inference based on students answering few items correctly. For standard 8.6.3, the assessment provides an opportunity to classify students at all four of the performance levels.

Table 32. Consensus item ratings from the Eastern regional meetings for CRTs in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Eastern districts (non-common)

	District 1-E					District 2-E					District 3-E				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.4	12	0	6	6	0	16	2	7	6	1	22	0	8	14	0
Standard 8.2.2	8	0	2	6	0	NR					22	0	14	6	2
Standard 8.3.2	13	0	3	10	0	30	0	8	22	0	23	4	8	11	0
Standard 8.4.1	60	10	15	27	8	27	0	5	16	6	27	0	6	20	1
Standard 8.5.2	40	8	7	17	8	40	1	13	20	6	26	3	11	10	2
Standard 8.6.3	5	0	3	2	0	10	0	6	4	0	26	8	12	5	1

As was the case for the grade 4 assessments, assessments from two districts were rated across at three meetings. In District 1-All, the assessments for only four standards provided sufficient information to make any student classifications. The assessments for standards 8.5.2 and 8.6.3 did not have enough assessment tasks at any performance level to make performance level classifications. The assessments for standards 8.1.4 and 8.3.2 permit classifying students as Proficient, Progressing, or Beginning. The assessments of standards 8.2.2 and 8.4.1 permit classifying students as either Progressing or Beginning. The assessments rated for District 2-All vary widely as to the number of assessment tasks provided to the students, ranging from as few as 4 items (standard 8.2.2) that permits no classifications, to a high of 92 items (standard 8.1.4) permitting classifications of Proficient, Progressing, and Beginning. The assessments for standards 8.5.2 and 8.6.3, like the assessment for standard 8.2.2 do not permit making any classification decisions. Students may be classified into all four levels using the assessment for standard 8.3.2 and into the three lowest levels using the assessment for standard 8.4.1.

Table 33. Consensus item ratings for CRTs in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Common districts.

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.4	29	0	5.7	23.7	0	92	4.7	29.7	56.7	1
Standard 8.2.2	16	0	3.3	9	3.7	4	0	4	0	0
Standard 8.3.2	20	0.7	7	12.3	0	42	9.3	15	15.7	1.7
Standard 8.4.1	40	0	3.7	28.7	7.7	26	1.3	9.7	15	0
Standard 8.5.2	7	0.7	2.3	2	2	6	0	3.3	2	0.7
Standard 8.6.3	4	1	2	1	0	5	0	3	2	0

NRT-2 was rated in the Central regional meeting. This NRT has two subtests that contain items aligned to standard 8.2.2 and one subtest having items related to standard 8.5.2. The six items aligned with standard 8.5.2 are distributed to widely to provide enough measurement information to make any performance level classifications. The 25 items (across both subtests) aligned with standard 8.2.2 permit classification into Proficient, Progressing, and Beginning (by inference) categories.

Table 34. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Math Concepts and Problem Solving					
Standard 8.2.2	10	1	6	3	0
Standard 8.5.2	6	1	2	3	0
Math Computation					
Standard 8.2.2	15	0	9	3	3
Overall					
Standard 8.2.2	25	1	15	6	3
Standard 8.5.2	6	1	2	3	0

Table 35 below shows the data for the two districts' assessments that were rated uniquely in the Central regional meetings (District 1-C's assessment was not rated at Grade 8). In District 2-C, the assessment for standard 8.1.4 can be used to classify students as either Proficient or Below Proficient. No classification decisions can be made with the five items associated with standard 8.2.2. For standard 8.3.2, 8.4.1, and 8.6.3 students can be classified as Progressing or Beginning using the appropriate assessments. For standard 8.5.2, a classification of Above Beginning could be made for students who answer more than six items correctly.

In District 3-C, three standards had no assessments rated. The remaining three assessments permit classification into multiple performance levels. Specifically, the 15 items associated with standard 8.3.2 permit classification of students into Proficient, Progressing and Beginning (by inference) levels. The assessment for standard 8.4.1 permits only two levels of classification, Progressing and Beginning. Standard 8.6.3's assessment provides the opportunity to say that students are Above Progressing (for those who answer virtually items correctly), Progressing, or Beginning (by inference).

Table 35. Consensus item ratings from the Central regional meetings for CRTs in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Central districts (non-common)

	District 2-C					District 3-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.4	15	0	11	4	0	NR				
Standard 8.2.2	5	0	3	2	0	NR				
Standard 8.3.2	12	0	4	8	0	15	0	8	6	1
Standard 8.4.1	32	0	0	25	7	24	0	2	19	3
Standard 8.5.2	8	0	5	3	0	NR				
Standard 8.6.3	15	1	4	8	2	22	1	5	13	3

The is only one subtest that includes mathematics items associated with standards 9.2.2 and 8.5.2. There are 10 items aligned with each of these standards. The items for standard 8.2.2 are distributed in such a way that permits classifications of Above Progressing (for

those who answer almost all items correctly) or Below Proficient (for those who answer 6 or fewer items correctly). The 10 items associated with standard 8.5.2, permit classifying students as either Proficient or Below Proficient.

Table 36. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 8 for Standards 8.2.2 and 8.5.2.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Mathematics Subset					
Standard 8.2.2	10	3	5	2	0
Standard 8.5.2	10	1	6	3	0
Overall					
Standard 8.2.2	10	3	5	2	0
Standard 8.5.2	10	1	6	3	0

Two of the three non-common districts, tended to have at least ten or more items per standard. District 1-W tended to have the largest number of measurement opportunities (22 to 26 items) across all six standards, whereas, District W-3, had only 4 to 8 items per standard.

District 1-W's assessments permitted classifications at the three lowest performance levels for five of the six standards of interest in this study, standards 8.1.4, 8.2.2, 8.3.2, and 8.5.2. The assessment for standard 8.4.1 had items that limited classification to only Progressing and Beginning performance levels. Standard 8.6.3's assessment provided sufficient breadth to assess and assign students to all four performance levels.

Table 37. Consensus item ratings from the Western regional meetings for CRTs in Mathematics for Grade 8 for Standards 8.1.4, 8.2.2, 8.3.2, 8.4.1, 8.5.2, and 8.6.3 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 8.1.4	22	0	7	15	0	10	0	4	6	0	6	0	4	2	0
Standard 8.2.2	22	0	13	6	3	15	0	1	8	6	4	1	3	0	0
Standard 8.3.2	23	0	11	11	1	14	0	6	4	0	8	0	6	2	0
Standard 8.4.1	27	0	2	21	4	14	0	0	9	5	8	0	3	5	0
Standard 8.5.2	26	4	5	13	4	5	0	2	2	1	7	1	3	3	0
Standard 8.6.3	26	14	2	9	1	12	3	2	7	0	10	2	7	1	0

Mathematics High School

The results of teachers' ratings of NRT-1 for high school standards 12.2.1 and 12.5.1 are shown in Table 38. Items in two subtests are aligned with standard 12.2.1 resulting in a total of 28 aligned items. These 28 items were judged to be evenly divided between the Proficient and Progressing categories, thus permitting classifying students as either Proficient, Progressing, or Beginning (by inference). There were only seven items aligned

with standard 12.5.1 and these items ranged across three performance levels without sufficient focus to make any classification decisions.

Table 38. Consensus item ratings from the Eastern regional meeting for NRT-1 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Mathematics: Concepts and Problem Solving					
Standard 12.2.1	17	1	9	7	0
Standard 12.5.1	7	0	3	1	3
Computation					
Standard 12.2.1	11	0	4	7	0
Overall					
Standard 12.2.1	28	1	13	14	0
Standard 12.5.1	7	0	3	1	3

The three districts that were evaluated as "non-common" in the Eastern regional meeting varied considerably in their coverage of the standards and their utility in assigning students to performance levels. In District 1-E the number of items ranged from a low of seven to a high of 17. The two standards for which there are only seven assessment tasks, standards 12.5.1 and 12.6.3, permit no classification into performance levels. The 11 items that are designed to assess standard 12.4.5 are also too diverse in difficulty to permit making any performance level decisions. Students who take the assessments for standards 12.1.2 and 12.2.1 can be classified as being either Progressing or Beginning, whereas the assessment for standard 12.3.1 permits only an inference about Beginning students.

District 2-E had two assessments that were not rated (for standards 12.3.1 and 12.5.1) and two other standards (12.1.2 and 12.4.5) that had too few items to permit any student performance level classification. For standard 12.2.1 the assessment permits three classifications of student performance, Advanced, Proficient, or Below Proficient (by inference). The assessment for standard 12.6.3 permits classifying students as either Proficient or Below Proficient.

All except one assessment from District 3-E permitted at least one level of performance classification. The one exception is the assessment for standard 12.5.1. The assessment for standard 12.3.1 permits assigning students to only the Beginning level of performance. Students can be assigned to either the Proficient, Progressing, or Beginning performance levels based on the assessments for standards 12.1.2 and 12.2.1, whereas for standards 12.4.5 and 12.6.3 students can be classified only as Proficient or Below Proficient.

Table 39. Consensus item ratings from the Eastern regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3 for Eastern districts (non-common).

	District 1-E					District 2-E					District 3-E				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.2	15	0	0	9	6	5	0	2	2	1	19	2	6	10	1
Standard 12.2.1	17	0	0	11	6	20	7	11	2	0	16	0	7	8	1
Standard 12.3.1	9	0	0	4	5	NR					14	0	2	3	9
Standard 12.4.5	11	1	2	4	4	3	0	0	3	0	16	3	11	2	0
Standard 12.5.1	7	2	3	2	0	NR					5	0	1	3	1
Standard 12.6.3	7	1	2	1	3	10	2	8	0	0	12	2	9	1	0

In District 1-All, at least two levels of classification can be made for all standards except standard 12.5.1, for which no classifications are possible. The assessments for standards 12.1.2 and 12.3.1 permit assigning students to either Progressing or Beginning performance levels. For standards 12.2.1, 12.4.5, and 12.6.3, the assessments permit classifications of Proficient or Below Proficient. None of the assessments permit a classification of Advanced.

Three standards (standards 12.2.1, 12.3.1, and 12.5.1) were not rated for District 2-All. Of the remaining standards, the assessment for standard 12.1.2 permits classifications of Advanced or Below Advanced. The assessment for standard 12.6.3 also permits two levels of classification, either Proficient or Not Proficient. Only the assessment for standard 12.4.5 permits three levels of classification, Proficient, Progressing, and Beginning (by inference).

Table 40. Consensus item ratings from the Eastern regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3 for Common districts.

	District 1-All					District 2-All				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.2	17	0	1.3	12	3.7	14	10.7	3.3	0	0
Standard 12.2.1	15	3.3	8.3	3.3	0	NR				
Standard 12.3.1	12	0	1.3	8	2.7	NR				
Standard 12.4.5	11	2	6	3	0	34	2.3	13	17.3	1.3
Standard 12.5.1	7	2	2.7	1.7	1	NR				
Standard 12.6.3	16	1.7	14	0.3	0	11	2.3	8.3	0.3	0

NRT-2 has only one subtest that includes items that were aligned with standards 12.2.1 and 12.5.1 and there are 10 and six items, respectively, associated with these two standards. The preponderance of items associated with standard 12.2.1 are at the Progressing level, thus permitting classification of students at that level and at the Beginning level (by inference). The six items aligned with standard 12.5.1 are distributed across the Beginning, Progressing, and Proficient performance levels prohibiting any determination of student performance levels.

Table 41. Consensus item ratings from the Central regional meeting for NRT-2 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Mathematics					
Standard 12.2.1	10	1	1	7	1
Standard 12.5.1	6	0	2	2	2
Overall					
Standard 12.2.1	10	1	1	7	1
Standard 12.5.1	6	0	2	2	2

Table 42 contains the data for the two non-common districts that provided mathematics assessments for review (District 3-C's assessments were not rated at Grade 12). District 1-C has only two assessments that permit student performance level classification. These are the assessments for standard 12.1.2 and 12.3.1 both of which permit classification of Progressing or Beginning. Two of the assessments, for standards 12.2.1 and 12.5.1 were not rated, and the remaining two assessments had too few assessment tasks to provide enough information to make a decision (however, the five items for standard 12.4.5 are all at the Proficient level suggesting a cautious performance level inference).

In District 2-C, students can be classified on standard 12.1.2 as Progressing or Below Progressing. For standard 12.2.1, the 26 items permit classification of students as Proficient, Progressing, or Beginning (by inference). The assessment for standard 12.4.5 permits classification as Proficient or Below Proficient (with a cautious inference at the Progressing level and, by inference, Beginning level). Assessments for the remaining standards, 12.3.1, 12.5.1, and 12.6.3, do not provide enough measurement information to make any performance level judgments.

Table 42. Consensus item ratings from the Central regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3 for Central districts (non-common).

	District 1-C					District 2-C				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.2	15	0	0	15	0	11	0	3	6	2
Standard 12.2.1	NR					26	0	11	15	0
Standard 12.3.1	12	0	1	7	4	6	0	0	3	3
Standard 12.4.5	5	0	5	0	0	11	2	4	5	0
Standard 12.5.1	NR					7	0	4	3	0
Standard 12.6.3	2	0	1	1	0	4	1	3	0	0

NRT-3 provides 29 items across two subtests that are aligned with standard 12.2.1. Most of these items are at the Progressing and Proficient performance levels, permitting classifications at those levels and at the Beginning level (by inference). The 16 items aligned with standard 12.5.1 are mostly at the Progressing level, permitting classification at the Progressing and Beginning performance.

Table 43. Consensus item ratings from the Western regional meeting for NRT-3 in Mathematics for Grade 12 for Standards 12.2.1 and 12.5.1.

Subtest	Items	Advanced	Proficient	Progressing	Beginning
Mathematics					
Standard 12.2.1	13	2	3	6	2
Standard 12.5.1	8	0	0	2	6
Mathematics Computation					
Standard 12.2.1	16	0	4	12	0
Overall					
Standard 12.2.1	29	2	7	18	2
Standard 12.5.1	8	0	0	2	6

The reviews of the assessments from the three non-common districts reviewed in the Western regional meeting are shown in Table 44. Only the assessments for District 1-W provide sufficient information for making any performance level decisions. In District 1-W, for standard 12.1.2 students may be classified as Proficient, Progressing or Beginning (by inference). For standard 12.2.1 students may be classified as Proficient or Below Proficient, whereas for standard 12.3.1 students may be assigned to either the Progressing or Beginning levels. The assessments for standards 12.4.5 and 12.6.3 permit classification of students into Proficient or Below Proficient categories. The assessment for standard 12.5.1 does not provide sufficient measurement information to make any determination of student performance levels.

In Districts 2-W and 3-W, there are not enough items at any single performance level to make classification decisions. However, note that for standard 12.3.1 in District 2-W a student who obtains most of the available points could be classified as being at least Progressing. A similar situation exists for this standard in District 3-W where a student who obtained all the available points (seven points), could be considered to be at least Progressing.

Table 44. Consensus item ratings from the Western regional meetings for CRTs in Mathematics for High School for Standards 12.1.2, 12.2.1, 12.3.1, 12.4.5, 12.5.1, and 12.6.3 for Western districts (non-common)

	District 1-W					District 2-W					District 3-W				
Standard	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**	**Items**	**A**	**Prof**	**Prog**	**B**
Standard 12.1.2	19	4	5	10	0	8	0	3	2	3	5	0	1	4	0
Standard 12.2.1	16	3	8	5	0	6	0	1	4	1	5	0	5	0	0
Standard 12.3.1	14	0	3	4	7	10	0	3	4	3	7	0	3	3	1
Standard 12.4.5	16	3	12	1	0	5	0	5	0	0	5	1	0	4	0
Standard 12.5.1	5	0	1	0	4	5	0	0	0	5	5	0	3	2	0
Standard 12.6.3	12	2	9	1	0	5	0	5	0	0	5	0	5	0	0

EVALUATION

At each of the three regional meetings there was an evaluation conducted to discern if there were any specific problems that needed to be addressed at future meetings of this type. Teachers completed the evaluation at the end of their ratings. All evaluations were anonymous.

The evaluation for was the same for all sessions, except the location was specific to the particular meeting and the subject area was specified for each group. Thus, there were six evaluation forms, each of which had the same evaluation items, but different headings. An example of an evaluation form is shown in Appendix C.

The evaluation form consists of 10 items across four parts. Part 1 provides participants an opportunity to rate the Orientation to the study. There are five items in this section, three items using a 6-point scale to rate the Study Overview, the Overview of the Performance Descriptions and Understanding the Role of the Participant. The remaining two items use a 3-point scale to judge if the amount of time for the orientation and for completing the practice test was too little, about right, or too much. Part 2 of the evaluation consisted of two items using a 4-point scale to rate the participants' comfort in judging the NRT and to assess if the participants' felt there was sufficient time allotted to this task. Part 3 was essentially the same as Part 2, except the focus was on the CRTs that were evaluated. Part 4 was an overall evaluation that provided two opportunities to rate the activity overall and in terms of the organization of the activity. This section also included an open-ended question asking for comments. Also included in Appendix C are all the comments from all the teachers who made comments on the evaluation form.

In essence the evaluation has three components. The first component is related to the quality of the experience, which is reflected in Parts 1 and 4. The second component is the adequacy of the time allotted to the activities, which is reflected by selected items in Parts 1, 2, and 3. The third component demonstrates the respondent's comfort with the results of the process. This is perhaps the most critical of the components of the evaluation because even if the experience was considered of high quality and the time allotment was adequate, if the participants do not feel good about their ratings, then little faith can be put into the finding. The evaluations from each of the regional meetings by content area are described below in terms of these three evaluation components.

Eastern Regional Meeting Evaluation – Reading

At the Eastern regional meeting in Lincoln, the teachers were very positive about the quality of their experience. All ratings were at the high end of the scale. The only problem areas were related to the time allotted to the rating task, by the high school teachers. Several of the high school teachers felt that not enough time was provided for the rating process.

In terms of their comfort ratings, teachers at all levels felt very comfortable with their rating decisions. All the ratings were above 3.5 on a 4-point scale.

Comments

Grade 4 teachers were, in general made positive comments about the process and the activity. Illustrative comments are shown below.

I felt that this experience has provided me with some necessary skills that I could use for my own purpose and to share with others.

Had difficulty matching items with rubrics (on some items).

One of the grade 8 teachers made a suggestion for how the activity might have been organized. The other teacher's comment was simply "Great".

The high school teachers had comments related to the room and conditions. In general they felt it was too cold and that the houseflies should have been under control. One teacher indicated that more time was needed; otherwise the comments were positive about the activity.

Table 45. Evaluation summary for teachers who rated reading assessments at the Eastern regional meeting.

Item	Grade 4		Grade 8		Grade 12	
	N	Rating	N	Rating	N	Rating
Part 1 Orientation (6-pt)	8	5.88	7	5.29	7	5.57
Part 1 Perf. Desc. (6-pt)	8	5.63	7	5.14	7	5.29
Part 1 Role (6-pt.)	8	5.25	7	5.14	7	5.29
Part 1 Orient. Time (3-pt.)	8	2.00	7	2.00	7	2.14
Part 1 Orient. Prac. (3-pt.)	8	2.00	7	2.00	7	2.43
Part 2 Comfort NRT (4-pt.)	8	3.75	7	3.57	7	3.43
Part 2 Time NRT (4-pt.)	8	3.25	7	3.29	7	2.71
Part 3 Comfort CRTs (4-pt.)	8	3.75	7	3.57	7	3.57
Part 3 Time CRTs (4-pt.)	8	3.25	7	3.14	7	2.43
Part 4 Overall (4-pt.)	8	3.38	7	3.14	7	3.14
Part 4 Organization (4-pt.)	8	3.88	7	3.29	7	3.29
Part 4 Comments	8	5 comments	7	2 comment	7	5 comments

Central Regional Meeting Evaluation – Reading

The teachers who participated in the Central regional meeting were generally positive about the quality of their experience. All ratings were at the high end of the scale, with only one rating related to the overview of the process being rated slightly lower (4.88 on the 6-point scale). One teacher in the grade 8 group and one high school teacher felt that not enough time was provided for the rating process.

In terms of their comfort ratings, teachers at all levels felt very comfortable with their rating decisions. All the ratings were above 3.4 on a 4-point scale.

Comments

At grades 4 and 8 only two of the comments were substantive. One fourth grade teacher recommended that the consensus form be a different color. The eighth grade teacher's comment was also repeated by several of the high school teachers. Specifically, it was recommended that the assessments be better aligned to the standards. Several high school teachers also commented on the poor alignment of the assessments to the standards. The

other high school teachers' comments were positive about the experience. The following summarizes those comments:

> It has been quite interesting to me to review and evaluate the assessments used by our schools. I was amazed at how similar the assessments were. I'm curious about the sources of each assessment (Internet, ESUs, etc.). How many were developed by the local schools? Were they duplicated and then sent out to other schools? I enjoyed the process of evaluation and discussion of each question on the assessments. It's always intriguing to hear other teacher's justification for a response. I thought Jim, Chad, & Renee did a great job with the workshop and I fount that their instructions were very explicit. It was well organized.

Table 46. Evaluation summary for teachers who rated reading assessments at the Central regional meeting.

Item	Grade 4		Grade 8		Grade 12	
	N	Rating	N	Rating	N	Rating
Part 1 Orientation (6-pt)	8	4.88	5	5.40	8	5.50
Part 1 Perf. Desc. (6-pt)	8	5.25	5	5.00	8	5.63
Part 1 Role (6-pt.)	8	5.50	5	5.00	8	5.88
Part 1 Orient. Time (3-pt.)	8	2.00	5	2.00	8	2.13
Part 1 Orient. Prac. (3-pt.)	8	2.00	5	2.20	8	2.13
Part 2 Comfort NRT (4-pt.)	8	3.88	5	3.60	8	3.75
Part 2 Time NRT (4-pt.)	8	3.25	5	3.00	8	3.13
Part 3 Comfort CRTs (4-pt.)	8	4.00	5	3.40	8	3.50
Part 3 Time CRTs (4-pt.)	8	3.25	5	2.80	8	3.25
Part 4 Overall (4-pt.)	8	3.13	5	2.80	8	3.13
Part 4 Organization (4-pt.)	8	3.38	5	3.00	8	3.38
Part 4 Comments	8	3 comments	5	2 comments	8	6 comments

Western Regional Meeting Evaluation – Reading

At the Western regional meeting in Lincoln, the teachers were very positive about the quality of their experience. Grade 8 teachers rated the orientation session lower than grade 4 or high school teachers. All ratings were at the high end of the scale. One high school teacher indicated that the orientation took too long. The only problem areas were related to the time allotted to the rating task, by the grade 9 and the high school teachers. Several of the teachers felt that not enough time was provided for the rating process.

Even though some teachers wanted more time to make their ratings, the teachers at all levels felt very comfortable with their rating decisions. All the ratings were above 3.5 on a 4-point scale.

Comments

All of the comments from the fourth grade teachers were extremely positive. The following comment is longer than most, but summarizes all the substantive comments.

> This was a very well organized study. The directions were clear and the purpose became even more clear as we got into this. I have been really frustrated with the

> lack of direction and explanation about tests up to this point and this really helped. I feel that while the purpose was to help with the study, I really benefited from this as well. It was a real eye opener and the collaboration was great! Thank you!

The Grade 8 teachers were more ambivalent than the Grade 4 teachers. They noted that it was useful and helpful to collaborate with peers, but that the ratings may not be as reliable as would be hoped. This latter feeling is expanded in the comment below.

> I've participated twice, Lincoln and Scottsbluff. I've enjoyed the work BUT, I am not at all confident that our ratings are reliable. Group dynamics affect the consensus too much. In Lincoln we had a "battleaxe:" and her opinion won out because people were too tired or afraid to argue with her. In Scottsbluff we had a very knowledgeable group but very congenial and eager to reach consensus. We "met in the middle" a fair number of times. The benefits of participation were a thorough discussion of standards and items by the group. I think the ratings will vary by group. I'd like to know the results. I do enjoy working for Buros because I learn so much. I rated the overall evaluation unsuccessful. Define success. If it was to reliably classify items I don't think our results are reliable.

High school teachers provided some insights to the potential for unreliability. One of the teachers indicated that the "green sheets" (performance level definitions) were not very helpful, so they relied on their classroom experience. As was the case in the Central regional meeting, one teacher reflected on the poor alignment of the assessment tasks to thee standards. Thus, although there were positive comments, several of the comments indicated potential problems that will need to be corrected in later studies.

Table 47. Evaluation summary for teachers who rated reading assessments at the Western regional meeting.

Item	Grade 4		Grade 8		Grade 12	
	N	Rating	N	Rating	N	Rating
Part 1 Orientation (6-pt)	9	5.56	7	4.43		5.25
Part 1 Perf. Desc. (6-pt)	9	5.67	7	4.43	8	5.63
Part 1 Role (6-pt.)	9	5.78	7	4.29	8	5.50
Part 1 Orient. Time (3-pt.)	9	2.00	7	2.00	8	1.88
Part 1 Orient. Prac. (3-pt.)	9	2.00	7	2.14	8	2.50
Part 2 Comfort NRT (4-pt.)	9	3.89	7	3.57	8	3.63
Part 2 Time NRT (4-pt.)	9	3.11	7	3.29	8	2.75
Part 3 Comfort CRTs (4-pt.)	9	3.89	7	3.57	8	3.50
Part 3 Time CRTs (4-pt.)	9	3.11	7	2.86	8	2.38
Part 4 Overall (4-pt.)	9	3.67	7	3.14	8	3.38
Part 4 Organization (4-pt.)	9	3.67	7	3.71	8	3.50
Part 4 Comments	9	7 comments	7	5 comments	8	6 comments

Evaluation Mathematics

Eastern Regional Meeting Evaluation – Mathematics

At the Eastern regional meeting in Lincoln, the teachers were very positive about the quality of their experience. Although some of the ratings of the orientation were less than 5.0 on the 6-point scale, the ratings were all above 4.0. As was the case in the reading meetings, the principal problem area was related to the time allotted to the rating task, by the high school teachers. One grade 4 teacher indicated that too much time was devoted to the practice test, whereas several of the grade 8 and high school teachers felt that not enough time was provided for the rating process on the practice test Several of the grade 8 teachers also felt that not enough time was allocated for rating the CRTs. This was not the case for the grade 4 and high school teachers.

In terms of their comfort ratings, teachers at all levels felt very comfortable with their rating decisions. All the ratings were above 3.6 on a 4-point scale.

Comments

Just over half the 11 fourth grade teachers made comments. These comments were mixed in there focus and in their evaluation of the process. Several teachers commented about the need for more of a stipend to participate, not just substitute pay. There was also a comment about the rating form, the rubric, and the availability of materials (only 6 teachers were expected and 11 showed up, so there were not enough copies of the NRT for everyone). Some illustrative comments include:

> Taught me so much! Some compensation even though sub pay for district. I worked non-contract hours to get ready to have a sub! (Just a thought).
>
> I felt frustrated at times – especially when had to share materials because there weren't enough individual copies.
>
> Use assessments that are well-done (CRT) for practice examples.
>
> Align practice items recording sheet going the same direction as definitions of student performance sheets for less confusion (e.g., beginning, progressing, proficient, advanced).

The eighth grade teachers had no comments and only two of the high school teachers made comments. These two comments were both positive.

> This provided good feedback for our own district.
>
> I was enlightened.

Table 48. Evaluation summary for teachers who rated mathematics assessments at the Eastern regional meeting.

Item	Grade 4		Grade 8		Grade 12	
	N	Rating	N	Rating	N	Rating
Part 1 Orientation (6-pt)	11	4.73	5	4.60	6	5.83
Part 1 Perf. Desc. (6-pt)	11	4.82	5	5.00	6	5.67
Part 1 Role (6-pt.)	11	4.73	5	5.00	6	5.83
Part 1 Orient. Time (3-pt.)	11	2.18	5	2.00	6	2.00
Part 1 Orient. Prac. (3-pt.)	11	1.91	5	2.40	6	2.17
Part 2 Comfort NRT (4-pt.)	11	3.70	5	3.80	6	4.00
Part 2 Time NRT (4-pt.)	11	3.30	5	3.40	6	3.17
Part 3 Comfort CRTs (4-pt.)	11	3.70	5	3.60	6	4.00
Part 3 Time CRTs (4-pt.)	11	3.20	5	2.40	6	3.33
Part 4 Overall (4-pt.)	11	3.20	5	3.40	6	3.33
Part 4 Organization (4-pt.)	11	3.27	5	3.40	6	3.83
Part 4 Comments	11	6 comments	5	0 comments	6	2 comments

Central Regional Meeting Evaluation – Mathematics

Teachers attending the Kearney meeting were generally positive about the quality of their experience. All ratings were at the high end of the scale. One grade 4 teacher wanted more time for the orientation and for the practice test. Otherwise all ratings tended to indicate that the timing was about right for the rating process.

In terms of their comfort ratings, teachers at all levels felt very comfortable with their rating decisions. All the ratings were above 3.6 on a 4-point scale. Of the six comfort ratings, three were 4.0 on the 4-point scale.

Comments

The comments from the mathematics teachers were virtually all positive except for two teachers who wanted better communications. However, these teachers did not expand on the comment other than to say that better communications were needed. The remaining teachers at all three grades indicated that it was a positive growth experience for them and that the process would be helpful in their districts.

Table 49. Evaluation summary for teachers who rated mathematics assessments at the Western regional meeting.

Item	Grade 4		Grade 8		Grade 12	
	N	Rating	N	Rating	N	Rating
Part 1 Orientation (6-pt)	5	5.20	5	5.50	6	4.83
Part 1 Perf. Desc. (6-pt)	5	5.25	5	5.00	6	4.83
Part 1 Role (6-pt.)	5	4.75	5	5.40	6	4.83
Part 1 Orient. Time (3-pt.)	5	2.20	5	2.00	6	2.00
Part 1 Orient. Prac. (3-pt.)	5	2.20	5	2.20	6	2.00
Part 2 Comfort NRT (4-pt.)	5	3.60	5	4.00	6	4.00
Part 2 Time NRT (4-pt.)	5	3.00	5	3.20	6	3.33
Part 3 Comfort CRTs (4-pt.)	5	3.60	5	3.60	6	4.00
Part 3 Time CRTs (4-pt.)	5	3.20	5	3.20	6	3.33
Part 4 Overall (4-pt.)	5	3.60	5	3.20	6	3.67
Part 4 Organization (4-pt.)	5	3.20	5	3.60	6	3.67
Part 4 Comments	5	3 comments	5	2 comments	6	3 comments

Western Regional Meeting Evaluation – Mathematics

At the Western regional meeting in Scottsbluff, the teachers were very positive about the quality of their experience. All ratings were at the high end of the scale. Similarly, the ratings regarding time were also high, indicating that the timing was about right for conducting the rating tasks.

In terms of their comfort ratings, teachers at all teachers felt very comfortable with their rating decisions. Only two of all participating teachers rated their comfort as a 4 on a 4-point scale. The two teachers, who did not provide a rating of four, gave their comfort rating as a three.

Comments

Virtually all the comments from all three groups of teachers were of a positive nature. There were a couple of suggestions for future experiences. Specifically, that the order of the performance level definitions on the rubric be the same as the order on the rating forms and one teacher suggested coming to consensus as a group on each assessment task rather than doing all of the tasks related to a standard, then coming back to reach consensus.

Table 50. Evaluation summary for teachers who rated mathematics assessments at the Western regional meeting.

Item	Grade 4		Grade 8	
	N	Rating	N	Rating
Part 1 Orientation (6-pt)	4	4.50	7	5.00
Part 1 Perf. Desc. (6-pt)	4	4.50	7	4.86
Part 1 Role (6-pt.)	4	4.50	7	4.86
Part 1 Orient. Time (3-pt.)	4	2.00	7	2.14
Part 1 Orient. Prac. (3-pt.)	4	2.00	7	2.14
Part 2 Comfort NRT (4-pt.)	3	4.00	7	3.86
Part 2 Time NRT (4-pt.)	3	3.67	7	3.29
Part 3 Comfort CRTs (4-pt.)	3	4.00	7	3.86
Part 3 Time CRTs (4-pt.)	3	4.00	7	3.29
Part 4 Overall (4-pt.)	3	3.33	7	3.29
Part 4 Organization (4-pt.)	4	3.75	7	3.29
Part 4 Comments	4	3 comments	7	2 comments

CONSISTENCY STUDY RESULTS

The principal objective associated with conducting an examination of the ratings for the same assessments across all three sites was to investigate the extent that ratings were site independent. Specifically, given that all teachers at all sites received essentially the same training experience and used the same definitions for the performance levels, it was of interest to see the extent that the teachers would apply these definitions in a consistent way across sites.

Two analysis strategies were employed for this component of the study. The first was to examine the extent that ratings of item within an assessment were similar. For example, the ratings for all the items in the assessment for reading standard 4.1.1 in District 1-All were rated at the Progressing level at the Eastern regional meeting, were they all also rated as Progressing at the other two meetings? This analysis represents the discrepancy among ratings and is reported for the assessment as a whole, rather than for each assessment item/task.

The second analysis, which is similar to the first, but is more practical, examined the extent that different interpretations of the results would occur from the data across different sites. For example, if Eastern regional ratings of the reading assessment for standard 4.1.1 permitted classification of students into Progressing and Beginning categories, did the ratings in the other two regions permit the same classification decisions? For assessments that have many items, there might be many instances when item discrepancies occurred, but these discrepancies did not have an impact on the classification decisions.

Because of space limitations, only one district (district 1-All) is illustrated and discussed for reading at grade 4 and one district (District 2-All) is illustrated for grade 8 mathematics. The results across all grades, districts, and content areas are similar to those reported below. The tables associated with the remaining districts, grades, and content areas are shown in Appendix B.

Reading – Grade 4, District 1-All

Table 51 below shows the results for District 1-All in Reading for standard 4.1.1 the greatest discrepancy across regions was at the Progressing and Beginning levels. Both Central and Western regions tended to be in agreement with each other as compared to the ratings in the Eastern Regional meeting. There was some discrepancy in the number of items classified as Advanced by teachers in the Western region as compared teachers in the Eastern and Central regions.

In terms of classification decisions, the same classification would be made based on the Central and Western results. Specifically, students could be classified as Proficient, Progressing, or Beginning (by inference). If the results from the Eastern region prevailed, student could be classified as Proficient or Below Beginning because students who answer fewer than 8 items correctly could be Progressing or Beginning.

Table 51. Ratings of assessment for Reading standard 4.1.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	17	1	7	4	4
Central	17	1	7	8	0
Western	17	3	6	7	0
Overall	17	1.67	6.67	6.33	1.33

The results associated with the assessments for standard 4.1.2 are similar to those of standard 4.1.1 where the teachers most in agreement are those from the Eastern and Central regional meetings.

Regarding the agreement in making classification decisions across the three meetings, those also follow the pattern associated with the assessment for standard 4.1.1. If the classification decision was to assign students as either Proficient or Below Proficient, then there is complete agreement across the three sets of teachers.

Table 52. Ratings of assessment for Reading standard 4.1.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	16	1	7	8	0
Central	16	1	7	4	4
Western	16	3	6	7	0
Overall	16	1.67	6.67	6.33	1.33

These results for standard 4.1.3 exactly duplicate those for standard 4.1.2.

Table 53. Ratings of assessment for Reading standard 4.1.3 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	16	1	7	8	0
Central	16	1	7	4	4
Western	16	3	6	7	0
Overall	16	1.67	6.67	6.33	1.33

As shown in Table 54, the teachers who attended the Eastern and Central again had, for standard 4.1.4 a different perception of what constituted a Proficient and a Progressing assessment task, than the teachers at the Western meeting. At all three meetings there was agreement that all the assessment tasks were at the same level, only the application of the definition of that level was inconsistent.

Clearly, the classification decisions would differ depending on which group of teachers made the decisions about performance level. If this assessment had been rated only in the Western regional meeting, students would be classified as Progressing or Beginning (by default). If, however, it had been reviewed only in either the Eastern or Central region, then students would have been classified as Proficient or Below Proficient. This is a substantial disagreement across the regional meetings. The basis for this discrepancy in the Western region is not clear.

Table 54. Ratings of assessment for Reading standard 4.1.4 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning

Eastern	9	0	9	0	0
Central	9	0	9	0	0
Western	9	0	0	9	0
Overall	9	0	6	3	0

As can be seen in Table 55, the results for standard 4.1.5 are patterned there is a slight disagreement between the Eastern teachers and the teachers in the Central and Western meetings. This slight disagreement for two items did not have an impact on the overall interpretation across all three meetings. Specifically, each meeting would have resulted in an assessment that was judged to classify students as either Proficient or Below Proficient.

Table 55. Ratings of assessment for Reading standard 4.1.5 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	15	0	15	0	0
Central	15	0	13	2	0
Western	15	0	13	2	0
Overall	15	0	13.67	1.33	0

The lowest levels of agreement for District 1-All were on the assessment for standard 4.1.6. As shown in Table 56, the teachers at the Eastern and Central meetings agreed somewhat with their ratings of Advanced and Progressing, whereas the Central and Western teachers tended to agree on their assignment of assessment tasks in the Proficient category. Overall, the classification decisions would be essentially the same based on the ratings from the Eastern and Central teachers. These ratings would permit students to be classified as Progressing if more than six or seven points were awarded and Below Progressing if fewer than four or five points were awarded. However, the six or more correct would result in a classification of Proficient if the opinions of the teachers in the Western region prevailed, with the Below Proficient decision being made for students who answered fewer items correctly.

Table 56. Ratings of assessment for Reading standard 4.1.6 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	9	1	2	4	2
Central	9	1	4	5	0
Western	9	4	4	1	0
Overall	9	2	3.33	3.33	.67

Conclusions about common districts' reading assessments

As shown in Table 51 – 56 above, there is no consistent pattern of agreement or disagreement across the three sets of teachers at the regional meetings regarding their ratings of assessments across the six reading standards for District 1-All at grade 4. For the assessments related to certain standards, all teachers were in substantial agreement, whereas for others, lower levels of agreement were observed. There was a tendency for the Eastern and Central teachers to be in agreement in terms of their perceptions of what constituted Proficient work, but even that was not consistent across the six standards. It appears that the extent of agreement is a function of both the teachers' orientation and the particular assessment they are evaluating.

Regarding the extent that the same performance level classifications would be made, the results paralleled the degree to which assessment tasks were rated similarly. Specifically, for all six standards, even though there were some minor disagreements about specific items, the overall determination of student classifications would be the same across two out of the three regional meetings. There was no assessment that was rated such that the ratings from all three meetings would have resulted in the same classification decisions.

Mathematics – Grade 8, District 2-All

In Table 57, the results of the rating from each regional meeting for mathematics standard 8.1.4 are reported. As shown in the table, there are several discrepancies in ratings of assessment tasks across the performance levels. Specifically, at the Advanced level, the teachers in the Central and Western meetings tended to agree, whereas at both the Proficient and Progressing levels, the Eastern and Western teachers were in fairly close agreement. In these two categories, the Central region teachers felt that there were many fewer items at the Proficient level and more at the Progressing level than did either the Eastern or Western teachers. None of the assessment tasks were rated as Beginning in either the Eastern or Central regional meetings, whereas the Western teachers placed three items at that level.

In terms of making overall classification decisions the Eastern teachers would permit classifying students as Proficient, Progressing, or Beginning (by inference). Teachers at the Central and Western meetings identified enough assessment tasks at the Advanced level to permit placing students into all four levels. It is most helpful that there are a large number of assessment opportunities for students for this standard, thus permitting such robustness in classification, when there was disagreement among the teachers across the three regional meetings.

Table 57. Ratings of assessment for Mathematics standard 8.1.4 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	92	2	38	52	0
Central	92	6	18	68	0
Western	92	6	33	50	3
Overall	92	4.7	32	56.7	1

Unlike the assessment associated with standard 4.1.4, which had over 90 items, there are only 4 assessment opportunities for standard 8.2.2. There was complete agreement on the level of challenge for all four items across all regional meetings. Unfortunately, with only four items, there is not enough information to make any classification decisions.

Table 58. Ratings of assessment for Mathematics standard 8.2.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	4	0	4	0	0
Central	4	0	4	0	0
Western	4	0	4	0	0
Overall	4	0	4	0	0

There was substantial disagreement across the three regional meetings about the 22 items available for standard 8.3.2 for District 2-All. As can be seen in Table 59, the only agreement was between teachers in the Eastern and Central regional meetings regarding the assessment tasks at the Progressing Level. Although there was little overall agreement in terms of items, there was more agreement in making the determination of performance levels, but even those are not in total agreement. Arguments can be made that using the judgments of teachers in the Eastern and Western regional meetings; students can be classified into four performance levels. The ratings by the teachers in the Central regional meeting permit classification into only three levels (no Advanced classification is possible).

Table 59. Ratings of assessment for Mathematics standard 8.3.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	42	16	5	21	0
Central	42	2	15	20	5
Western	42	10	26	6	0
Overall	42	9.3	15	15.7	1.7

Teachers in the Eastern and Central regional meetings were in complete agreement on classifying items related to standard 8.4.1. Teachers in the Western regional meeting felt that, in general, the items were more likely to be at the Proficient level, than at the Progressing level. In making classification decisions, however, all three sets of teachers' ratings would result in permitting classifying students as Proficient, Progressing, or Beginning (by inference).

Table 60. Ratings of assessment for Mathematics standard 8.4.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning

Eastern	26	2	5	19	0
Central	26	2	5	19	0
Western	26	0	19	7	0
Overall	26	1.3	9.7	15	0

Although there are only six assessment tasks associated with standard 8.5.2, there was substantial disagreement across the three regions about classifying the items into performance categories. Specifically, in the Eastern regional meeting the teachers indicated all six items were at the Proficient level, the Central region teachers felt that only three items were at the Proficient level, and the Western region teachers identified only one item at that level. This disagreement is also reflected in terms of how students might be placed into performance levels. Specifically, if using the Eastern region teachers' judgment, students can be classified as Proficient or Below Proficient. If using the ratings of teacher in the Central region, no classification decisions are possible, but the Western region teachers provide classification of students as Progressing or Beginning (by inference)

Table 61. Ratings of assessment for Mathematics standard 8.5.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	6	0	6	0	0
Central	6	0	3	1	2
Western	6	0	1	5	0
Overall	6	0	3.3	2	.7

Unlike the assessment aligned with standard 8.5.2, there was complete agreement across the three regional meetings on the assessment associated with standard 8.6.3. No classification decisions can be made using these five assessment tasks.

Table 62. Ratings of assessment for Mathematics standard 8.6.3 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	5	0	3	2	0
Central	5	0	3	2	0
Western	5	0	3	2	0
Overall	5	0	3	2	0

Conclusions about common districts' mathematics assessments

The conclusions regarding the mathematics assessment are very similar to those reached regarding the reading assessment described above. As shown in Tables 57 – 62 above, there is no consistent pattern of agreement or disagreement across the three sets of teachers at the regional meetings regarding their ratings of assessments across the six mathematics standards for District 2-All at grade 8. For the assessments related to certain standards, all teachers were in substantial agreement, whereas for others, lower levels of agreement were observed. Unlike the reading assessments, there was no tendency for the Eastern and Central teachers to be in agreement in terms of their perceptions of what constituted Proficient work and also there was, for some standards, substantial agreement in the ratings across meetings. It appears that the extent of agreement is largely a function of both the teachers' orientation to the performance level definitions and the particular assessment they are evaluating.

Regarding the extent that the same performance level classifications would be made, the results suggest that there is higher agreement on this, than on exact classifications. Specifically, for three of the six standards, the overall determination of student classifications would be the same across all three of the regional meetings. For two of the remaining standards, teachers' ratings in two of the regional meetings would have resulted in making the same student classification decisions. For only one standard, the final classification decision would differ depending on which set of teachers made the performance level determination for the assessment.

CONCLUSIONS AND RECOMMENDATIONS

The conclusions and recommendations are divided into two parts. Part 1 relates to the process of making determinations about assessment tasks based on using a rubric consisting of performance level definitions. This includes not only judgments about the process but also about the consistency of those judgments across several different settings with different individuals. Part 2 of the conclusions focuses on the utility of the assessments that were judged for making performance level judgments about students. Both the NRTs and the CRTs that were evaluated by the teachers are discussed in terms of how well they can be used, in general to classify students into various performance levels. Recommendations are integrated into the text of the conclusions, where recommendations are appropriate.

Conclusions and recommendations Part 1 – The Process

The process involved developing performance level definitions (at an earlier meeting), obtaining assessments (NRTs and CRTs) for both practice and operational rating, preparing the materials (CRTs and rating forms) for evaluation, using the performance level definitions to evaluate the assessments. In addition, the process required the cooperation of districts to provide teachers and the willingness of the volunteered teachers to participate. The procedures were essentially the same for both reading and mathematics and at all the regional meetings, thus, the conclusions drawn will be generic, unless a specific issue is more relevant to a particular location or content area. Part of the process was to examine the consistency of judgments across meeting using common assessments. The conclusions and recommendations related to that aspect of the study are found at the end of that section of this report (see above).

Developing and using performance level definitions

In the design of this study, it was deemed important that there be overlap between the participants at the July meeting to develop the performance level definitions and the participants who rated the assessment tasks using those definitions. The justification for the overlap was that those who helped develop the definitions could assist in their interpretation in the operational phase. This outcome did not come to pass. The teachers who attended both meetings did not have sufficient recollection of the definitions or the discussions that had transpired in the development of those definitions. Moreover, they did not recall the rationale for making the decisions about performance levels for the assessment tasks used as practice.

Recommendation: The notion of overlapping participants may be sound, but the time span between the development of the definitions and the operational study needs to be much closer in time so that the participants who develop the definitions have a better opportunity to remember the discussions and rationale for the decisions made at the initial meeting.

Because the development of the performance level definitions was done at an earlier meeting and a report on that meeting has already been delivered, that process will be touched on only briefly in this report. Specifically, once the teachers at the July meetings in Lincoln had drafted the initial performance level definitions, these draft definitions were edited for consistency of language and format within each content area. In this

process, more care was needed to insure consistency. This was of particular importance in mathematics where several seemingly minor language issues caused interpretation problems. For example, regarding the "making change" component of the grade 4 standards, the progression from Beginning to Advanced had to do with the amount of money involved. The language of the performance levels was the same for Progressing, Proficient, and Advanced (for amounts of $5, $10, and unlimited, respectively), but was different for Beginning (for amounts rounded to the nearest $1.00). This caused some confusion among the participants who wanted consistency across all performance levels (e.g., the Beginning should have said for amounts of up to $1.00).

Recommendation: Before undertaking further studies using performance level definitions, the definitions should be reviewed and ambiguities and inconsistencies eliminated.

Another related issue regarding the performance level definitions is the determination of what element in the rubric takes precedence. Specifically, if there are multiple conditions that might apply to a particular assessment task, which conditions takes precedence. For example, if one element of the rubric suggests that a multi-step problem is at the Proficient level, and another element suggests that dealing with values under $5.00 is at the Progressing level and the assessment task requires two steps, but is for an amount under $5.00, which level should be selected? This happened most often in the mathematics content area, but not exclusively.

Recommendation: A rule of precedence must be developed when more than one standard might be applicable (either directly or indirectly) to an assessment task.

Similarly, there were some instances when the application of the performance level definition did not seem appropriate because the nature of the assessment was either so obvious or so complex that a different level was more suitable. For example, in the reading standards vocabulary is a relevant area of assessment. If the vocabulary is at the appropriate grade level, then the performance level may be Beginning or Progressing. However, in some assessments, the vocabulary that was tested was at the appropriate level, but the passage from which the vocabulary was drawn and the method used to assess the vocabulary was so complex that the teachers felt strongly that only Proficient or Advanced students would be able to gain any points.

Recommendation: Clarify the weight of the rubric and the weight of the assessment task when there appears to be a conflict. This may require expanding the performance level definitions to include not only content, but also skills that are needed to demonstrate content knowledge (i.e., how the content is assessed).

When there were ambiguities in the performance level definitions or when the performance level definitions did not cover the content of a standard, teachers often used their own perception of what a Beginning, Progressing, Proficient, or Advanced student was based on the students in their classes (and perhaps by the definitions used in their districts). In such cases the determination of performance level was not consistent either within the group or across groups. Decisions by teachers to abandon the agreed upon performance level definitions and revert to their individual and unique definitions may be partly responsible for the discrepancies across meetings in determining the performance levels of assessments that were rated across meetings.

Recommendation: Continually emphasize to the teachers that the decision about the performance level of an assessment task should not be based on their students but instead is based on the performance level definitions provided and on the rules of precedence and weights developed based on the above recommendations.

Obtaining and preparing NRTs and CRTs

Three NRTs were used in this study. Contact individuals at the publishers of these tests were asked to provide 6 copies of each test at each grade level (18 tests from each publisher). Two of the publishers responded almost immediately, whereas the third was reluctant to provide the needed tests. After some additional communications with the reluctant publisher and after contacting the sales representative (rather than the contact person in the research office) the tests were delivered. Once the NRTs were received, developing the rating forms was a very straightforward process. This is because the data from the previous alignment study were available and could be used to determine which test items were aligned with the standards used in this study. Also because all items were multiple choice, the development of the rating forms was also very straightforward.

Although not required by the publishers, tabs were applied to all subtests that were not being used in this study. This prevented teachers from casually looking through tests that were not part of the study (e.g., Social Studies). In addition, all teachers signed a Non Disclosure Agreement in which they agreed to not disclose information about any of the assessments they encountered in this study.

Obtaining CRTs was very challenging. This was in part due to confidentiality of the district assessments and in part due to the reluctance of the districts to be exposed in the event that their assessment might fall short of being of high quality.

Recommendation: In future studies, a better explanation of the outcomes of the process and reassurance that the results will be confidential (to other districts and the NDE) along with assurances that the results will be shared with the district, may help with getting cooperation.

Once assessments were received, much work was needed to get them ready for review. All assessment tasks had to be reviewed to ensure that they could be rated. In some districts, especially for the reading standards, the assessment task is defined only in vague terms that left substantive specific decision up to the teacher and the scoring rubric is so general that the assessment task could not be rated. For example, some assessments simply indicated that the teachers should identify a reading passage or selection, and have the students respond to a set of general questions about the passage (e.g., What was the plot?). The difficulty of the task would be a function of which passage or selection the teacher identified. Such assessment tasks were not rated because the variation from classroom to classroom could not be estimated.

Recommendation: Districts should be advised that more directions to teachers are needed. For example, rather than just leave open the selection of a reading passage, teachers should be provided a list of passages that are of similar difficulty.

In a similar vein, some of the mathematics assessments were constructed as dichotomous items, but the scoring system was based on an arbitrary rubric. For example, the assessment task was for the student to respond to a series of problems that could be

scored correct or incorrect (and a performance standard set based on the number correct), but the scoring was actually based on a rubric that classified students at various performance levels arbitrarily depending on the number of items they answered correctly. Buros staff recast such items so participants ignored the rubric and rated each of the assessment tasks (items) independently. This also happened in reading, but to a lesser extent. In a few instances, teachers rated assessments from their own districts and observed that the rubric was not part of the rating materials. Some explanation to these teachers was required, however, in all cases they immediately saw the logic of what we had done and agreed with the revision.

Recommendation: Additional professional development is needed to assist districts in understanding when rubrics are appropriate and when they are not. Moreover, districts need to know that using a rubric to define the performance levels may not be an appropriate procedure, because such a process bypasses the standard setting process and makes the performance level classification arbitrary.

Other preparation that was needed in advance of using many of the assessments was the deletion of the performance level definitions from the scoring rubrics. Many of the rubrics were constructed as 4-point scales, such that a score of one was defined and characterized as being "Beginning," a score of two was defined and characterized as being "Progressing," and so on. All such references to performance levels were eliminated because the definitions of the performance levels were to be based on the common definitions used in this study, not on the varying definitions across different school districts.

Recommendation: NDE should encourage all districts to develop performance level definitions and to apply these definitions independent of the rubrics used to score assessment tasks. If this is not done, then the assessment tasks should be evaluated by the district to make sure that the assessment tasks and the score levels accurately reflect the performance level definitions.

Because of the difficulty of obtaining CRTs to use in this study, the CRTs used to provide a practice experience for the participants did not provide enough illustrations of the various performance levels for each standard. This lack of breadth within the practice assessments meant that coming to group consensus about what types of assessment tasks illustrated the different performance levels within a standard was delayed until an operational example was found. This was often difficult and slowed down the learning process. An additional problem with the practice assessments was that they were not well aligned with the standards. This was a particular problem with the reading assessment, which involved a district that had its own standards and the alignment was done by Buros.

Recommendation: Practice tests should provide examples of all performance levels for all standards. They should be exemplary tests that represent quality assessments.

Obtaining participants

There are many competing activities that occur throughout the school year. Superintendents and Principals must make many decisions about the appropriateness and desirability of releasing teachers to participate in these various activities. It was very

difficult to obtain the cooperation of school administrators to release teachers to participate in this study. After substantial efforts by Renee Jacobson to recruit teachers, NDE was asked to intervene to obtain sufficient teacher to participate in this study. It is not clear how recruiting could be improved for future studies of this nature. It is interesting to note that many of the participants who made comments on the evaluation forms indicated that they had gained much from the process and that it had been a valuable learning experience for them. Such comments were also made verbally to project staff during and immediately following each of the regional meetings. Perhaps such testimonials could be used to help convince school administrators that the benefits of participation are sufficient to justify releasing their teachers.

Conclusions and recommendations Part 2 – Assessment Utility

Three commercially available norm-referenced tests (NRTs), each of which included one or more subtests related to each content area were examined to judge the extent that these tests could be used to classify students into four performance levels on a single standard in reading or on two standards in mathematics. The results were mixed.

In addition to these NRTs, a total of 11 CRTs were evaluated in reading and 10 in mathematics. The results across these 21 tests are very mixed in terms of their utility for classifying students in various performance categories.

This section of the report first discusses the NRTs, then the CRTs.

Reading NRTs

For the one reading standard (related to reading comprehension) examined at each of the three grade levels, the preponderance of test items were rated as being at the Proficient performance level. In general, the next most frequent performance level was the Progressing level. There were two exceptions to this both at grade 8 where two of the NRTs had more items at the Advanced level than they did at the Progressing level.

One test, NRT-1 has no items classified at the Beginning level on any of the subtest across the three grade levels. This NRT has insufficient items at the Advanced level to make confident classifications at the 4th grade. For the other two grade levels, this test could be used to classify students at all four performance levels (Beginning by inference).

NRT-2 has sufficient breadth of difficulty to classify students into the four performance levels for grades 4 and 8, but insufficient items at the high school level to classify students as Advanced.

Similarly, NRT-3 has enough breadth of item difficulty at grade 8 and high school to place students into the four performance levels (in some cases Beginning is by inference). However, there are not enough items at the Advanced level on the high school test to use it for student classification at that level.

In summary, each of these three tests can be used at two of the three grade levels to classify students into four performance levels. For the remaining grade, three performance levels can be used. Each test has only five items at the Advanced level on one of the tests, thus restricting its use at that level. If only two performance levels, Proficient or Below Proficient are desired, then all three tests can make that distinction.

It is known that many districts are using the NRTs to classify students into two performance levels using the 50%ile as the cut point between Proficient and Below Proficient. This is likely a reasonable cut point. It may be too high in some cases and too low in others, but overall it appears to be reasonable. It is also known that some districts are using NRTs to classify students into four performance levels where the cut point for Advanced is scores above the 75%ile and the cut point for Beginning is the 25%ile. The justification for using these tests in this way with these cut points is not justified by the data collected in this study.

Recommendation: Continue using the NRTs for classifying students as either Proficient and above or Progressing and below, but stop using NRTs for classifying students into four performance levels, especially using the arbitrary cut points of the 75%ile and 25%ile.

Mathematics NRTs

Each of the NRTs focused on two mathematics standards; one standard in strand 2 (Computation and Estimation) and one standard in strand 5 (Data Analysis, Probability and Statistical Concepts). Most of the NRTs had two subtests with items that related to these two standards, but some have only one and one has three subtests. The number and distribution of items varies considerably across the three NRTs for both standards.

For the strand 2 standard, NRT-1 has over 50 items across three subtests. The majority of these items are at the Proficient level and many are also at the Progressing level. There are five items at the Advanced level. Thus, decisions can be made about students at three levels, Proficient, Progressing, and Beginning (by inference as there are no items at this level). At grade 8, for this strand, there are only 14 items in a single subtest and 10 of those items are that the Proficient level, making it possible to classify students only as Proficient or Below Proficient. The high school test has enough items at the Proficient and Progressing levels to make three levels of classification (Beginning by default).

NRT-2 for the standard in strand 2, can be used to make classifications at all levels but Advanced at grades 4 and 8. At the high school level, of the 10 items, seven are the Progressing level, so decision about Progressing or Beginning (by inference) can be made.

Unlike NRT-1 and NRT-2, which provide no opportunity to classify students as Advanced on the strand 2 standard, NRT-3 does have enough items at the Advanced level at grade 4. It also has sufficient items at the Proficient and Progressing levels to make those classifications, and Beginning may be made by default. The picture changes at grade 8, however, because there are only enough items to make a decision for the strand 2 standard that students are Above Progressing or Below Proficient. At the high school level, all classification except Advanced can be made.

For the strand 5 standard all three NRTs have fewer items that for the strand 2 standard. This means that in general fewer classification decisions can be made. It is also noted that in general these items are often at the higher performance levels.

NRT-1 has sufficient items at grade 4 to classify a student as Proficient or Below Proficient. At grade 8, the decision of Proficient, Progressing, or Beginning (by inference) can be made. There are only seven items on the high school level test related

to the stand 5 standard and those items are distributed to broadly to make any classification decisions with confidence.

For NRT-2, the only grade at which any classifications can be made is grade 4, where students may be identified as Advanced, Proficient, or Below Proficient (by default). There are only 5 items at grade 8 and six items on the high school test for this standard and these items are distributed across the performance levels.

There are also limitations on the degree to which classifications can be made with NRT-3. At grades 4 and 8, there are sufficient items at the Proficient or Advance level to permit classifying students as either Proficient or Below Proficient. The high school test can only identify Beginning students (with 6 of 8 items at that level).

In summary, the NRTs in mathematics tend to be variable in their utility to classify students across the various performance levels. Overall, these tests do a better job for the strand 2 standard, in that they have more items related to this standard and these items tend to be distributed in their difficulty to permit, for most grades, making three performance level decisions. And when this level of decision making is not possible, then students can be classified as being Proficient or Below Proficient. For the standard related to strand 5, however, fewer decisions can be made about how well students are performing. Moreover, in only six of the nine test situations (three tests at three grade levels) can any classification decision be made. On the positive side, five of these six decisions can separate the Proficient from the Below Proficient student.

Recommendation: Districts that are using these NRTs to make classification decisions should be doing so with extreme caution. It would be useful to supplement these tests with well-constructed CRTs to be more comfortable in the classification of students into performance levels. The items related to the strand 5 standard represent only a minority of the items on the subtest on which theses items are found, thus using the 50%ile (or any other point) as the dividing point for Proficient or Below Proficient will probably result in many misclassifications.

Reading CRTs

A careful review of the results across all regional meetings and for each standard reveals that each standard is unique in terms of the utility of the districts' assessments that were used in this study. Although there is no consistent pattern of utility across the six reading standards, the review does show that for many of the grade 4 standards between 6 and 9 of the districts' assessments will permit classifying students as being Proficient or Below Proficient. Moreover, between 2 and 6 districts' assessments will permit classifications of Proficient, Progressing, and Beginning, and for the majority of standards at least one district's assessment will cover all four performance levels. Looking at these data in a slightly different way, we can consider that at grade 4 there were 10 district assessments for six standards or 60 opportunities to classify students. Of these 60 opportunities, there were 9 (15%) that provided for a classification of Advanced and 40 (67%) that provided classifications of Proficient or Below Proficient.

At grades 8 and high school, the coverage of the standards is similar to that at grade 4, but there are fewer standards that have that coverage. That is, fewer districts' assessments permit dividing students between Proficient and Below Proficient, fewer districts'

assessments permit the three lower levels of assignment, and fewer districts' assessments provide for Advanced students. Using the same analysis as above for grade 8 there were 11 district assessments across the six standards providing an opportunity for making 66 classifications. Of these 66 opportunities, only 4 (6%) provided a placement of Advanced and 33 (50%) permitted a classification of Proficient or Below Proficient. The CRTs used in the high school provided 60 classification opportunities with zero Advanced and only 15 (25%) permitting a classification of Proficient or Below Proficient.

Clearly, at all grade levels, there are relatively few standards for which students can be classified as being Advanced. At grade 4 many students can be classified as being Proficient or Below Proficient on most standards, and for some standards the classification can be slightly more precise. However, the picture is not as good at grades 8 and high school. At these grade levels, generally between one and six of the districts' assessments permit classification of Proficient or Below Proficient and fewer still provide more precise classification.

In summary, the district assessments that were evaluated in this study tended to be able to classify students at grade 4 as Proficient or Below Proficient, with some districts' assessments able to provide more precise classifications. At grades 8 and high school, there were many assessments that were not rated, and of those that were rated, generally less than half could be used to classify students as Proficient or Below Proficient and fewer still could make more precise assignments. Across all grades, there are relatively few assessment opportunities for Advanced students.

What is not known is how representative these assessments are of all the assessments in the state. If these assessments reflect the status of local assessments, then little faith can be placed in the districts' reports to the NDE of how many (or what percentage) of students are at each performance level. At best, collapsing the four categories into two (Proficient and Below Proficient) would provide a somewhat more accurate description of student performance levels, but even this is problematic for high school students. (Note that some of the assessments that were evaluated were already being redeveloped and some had already been replaced and were no longer in use.)

Recommendation: Districts should carefully review their assessments to insure that the assessment tasks, and where appropriate the assessment rubrics, provide opportunities for students at all performance levels to demonstrate their knowledge and skill relative to all standards. Districts should start with their assessments at the high school level.

Recommendation: The NDE should place little confidence in the districts' classifications of students based on these assessments. The reported data should be collapsed into only two categories Proficient and Below Proficient to obtain a more accurate representation of student performance levels.

Mathematics CRTs

The utility of the assessments in mathematics tends to parallel the results cited above in reading. For the grade 4 assessments there were 11 districts' assessments that were examined and there were six standards, providing 66 opportunities to determine student performance levels. Of these 66 opportunities, there were 15 (23%) that provided an

opportunity to classify students as Advanced and 38 (58%) that provided a classification of Proficient or Below Proficient.

Similarly, there were 60 measurement opportunities at grade 8 and five (8%) of these provided for an assignment of Advanced and 28 (46%) permitted a classification of Proficient or Below Proficient. Although there are slightly more opportunities for an Advanced classification in mathematics than in reading, the Proficient or Below Proficient decisions are slightly fewer than was the case for reading.

The high school assessments for mathematics also faired poorly, but they were slightly better than those in reading. Of the 60 measurement opportunities, there were two (3%) that permitted a classification of Advanced (compared to none in reading). Seventeen (28%) of the measurement opportunities permitted a classification of Proficient or Below Proficient as compared to the 25% in reading.

The variations between mathematics and reading are not sufficient to justify any different conclusions or recommendations. In essence these assessments did not provide sufficient opportunities, especially at grade 8 and high school, to give much confidence in the classification of students in more than two performance levels, and even two performance levels is risky for high school. The caution noted in the section above related to reading related to the extent to which these assessments are representative of those used across the state is appropriate here as well. However, based on the analysis presented for mathematics, the recommendations for reading assessments are repeated for mathematics.

Recommendation: Districts should carefully review their assessments to insure that the assessment tasks, and where appropriate the assessment rubrics, provide opportunities for students at all performance levels to demonstrate their knowledge and skill relative to all standards. Districts should start with their assessments at the high school level.

Recommendation: The NDE should place little confidence in the districts' classifications of students based on these assessments. The reported data should be collapsed into only two categories Proficient and Below Proficient to obtain a more accurate representation of student performance levels.

SUMMARY

This project had two phases and was undertaken during the period from March 2003 to April 2004. The data for this study were collected in two phases. Phase 1 was the determination of performance levels in reading and mathematics. This took place in July 2003 and a report was delivered in September 2003. This report covers the second phase of the project in which data were collected to determine the extent that 3 NRTs from different publishers and 21 CRTs (11 in reading and 10 in mathematics) could be used to classify students as performing at Advance, Proficient, Progressing, or Beginning levels on selected standards. In addition to this main focus (the determination of the sufficiency of the assessments to make performance level classifications) there was a second study that looked at the consistency of teachers across multiple sites to make the same performance level judgments about a sample of assessments.

A total of 66 teachers met at one of three locations around the state to judge the assessments in reading. An additional 52 teachers met at these same locations to judge the mathematics assessments. After an introduction to the study and undertaking a

practice activity, teachers spent approximately a day evaluating the tests at their grade level.

The results of this activity suggest that the NRTs have some utility for classifying students as either Proficient or Below Proficient on the one reading standard that was examined. Some additional, more precise classification is also possible, but such classifications should be made with caution because, in some cases multiple subtests are involved and the proportion of items that focus on the specific standard may be a relatively small proportion of the items on the subtest. This is even truer in mathematics where the utility of classifying student on two standards was examined. On the standard related to Computation and Estimation, the NRTs were comparable to the utility for reading. Similarly, the items were often across more than one subtest, so the proportion of the total items that focus on the standard is not known. However, for the standard related to Data Analysis, Probability and Statistical Concepts, there tended to be fewer items, all in one subtest and fewer opportunities to make performance level decisions.

The CRTs that were examined included some assessments that were locally developed and used only in the local district. Other assessments had been developed in consortia and were used in multiple districts. In general the utility of these assessments to classify students into multiple performance levels is mixed. At grade 4 in both reading and mathematics the assessments of about 67% of the standards provided classification into either Proficient or Below Proficient categories. This percentage declines rapidly to about 50% at grade 8 and about 25% at high school. If more precise classifications are desired (e.g., Advanced, Proficient, Progressing, Beginning) these percentages drop dramatically. The worst case is in attempting to classify students as Advanced. Across all grade levels, assessments and standards in reading (3 x 11 x 6) there were only 13 instances when a rating of Advanced could be made. The situation is only slightly better in mathematics where there are 22 such instances.

The study related to consistency of judgments across settings, suggested that some caution should be undertaken in interpreting the results related to the CRTs. Although it was often the case that there was agreement between two of the three sets of teachers, and occasional agreement among all three sets, there were also some assessments on which there was virtually no agreement. In short, there was no consistent pattern of agreement. That is, for some assessments the teachers in the Eastern and Central regions agreed with each other on their item classification judgments, for other assessments, the Western and Central teachers agreed, and for yet other assessments the Eastern and Western teachers agreed. These were not systematic for the same standards.

A number of recommendations were made. Some of these recommendations focus on the process (e.g., making it clearer how the performance level definitions should be used). Other recommendations focus on the results, particularly on the utility of the assessments to be used for making performance level classifications.

Appendix A Illustrative rating forms for reading and mathematics for NRT and CRT

Illustrative Rating Form for Practice Items

Reading Grade 4

Passage 1 Moving Game

Item #	Answered Correctly by about 67% of students[5] Advanced	Proficient	Progressing	Beginning	Standard
1	_____	_____	_____	_____	4.1.3
2	_____	_____	_____	_____	4.1.3
3	_____	_____	_____	_____	4.1.3
4	_____	_____	_____	_____	4.1.3
5	_____	_____	_____	_____	4.1.3
6	_____	_____	_____	_____	4.1.2
7	_____	_____	_____	_____	4.1.2
8	_____	_____	_____	_____	4.1.1
9	_____	_____	_____	_____	4.1.1
10	_____	_____	_____	_____	4.1.2
11	_____	_____	_____	_____	4.1.2
12 (5 points)	_____	_____	_____	_____	4.1.3
13 (5 points)	_____	_____	_____	_____	4.1.3
14	N/A				
15	N/A				
16	N/A				

Part IV

Passage 2 Excuses, Excuses

17 a	_____	_____	_____	_____	4.1.3
17 b	_____	_____	_____	_____	4.1.3
17 c	_____	_____	_____	_____	4.1.3
17 d	_____	_____	_____	_____	4.1.3
17 e	_____	_____	_____	_____	4.1.3
18 a	_____	_____	_____	_____	4.1.4
18 b	_____	_____	_____	_____	4.1.4
18 c	_____	_____	_____	_____	4.1.4
18 d	_____	_____	_____	_____	4.1.4
18 e	_____	_____	_____	_____	4.1.4
18 f	_____	_____	_____	_____	4.1.4
19	_____	_____	_____	_____	4.1.3

[5] Check only the **highest** performance level for MC items or items with "right/wrong" scoring. For tasks or items with a rubric (more points than just right/wrong scoring), check all levels for which scoring is possible.

Illustrative Rating Form for Practice Items

Mathematics Grade 8

Standard 8.1.4 (Benchmarks 3, 7, 12)

Answered Correctly by about 67% of students[6]

Item #	Advanced	Proficient	Progressing	Beginning
Top				
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____
4	_____	_____	_____	_____
7	_____	_____	_____	_____
8	_____	_____	_____	_____
Bottom				
2	_____	_____	_____	_____
4	_____	_____	_____	_____
5	_____	_____	_____	_____
7	_____	_____	_____	_____

Standard 8.2.2 (Vacation Project)

Answered Correctly by about 67% of students

Item #	Advanced	Proficient	Progressing	Beginning
Rubric scoring				
Travel Costs	_____	_____	_____	_____
Hotel Costs	_____	_____	_____	_____
Meal Costs	_____	_____	_____	_____
Cost Accounting	_____	_____	_____	_____

Standard 8.3.2 (Aqueduct Project)

Answered Correctly by about 67% of students

Item #	Advanced	Proficient	Progressing	Beginning
Rubric scoring				
Design	_____	_____	_____	_____

Standard 8.4.1 (Aqueduct Project)

Answered Correctly by about 67% of students

Item #	Advanced	Proficient	Progressing	Beginning
Rubric scoring				
Shapes	_____	_____	_____	_____

[6] Check only the **highest** performance level for MC items or items with "right/wrong" scoring. For tasks or items with a rubric (more points than just right/wrong scoring), check all levels for which scoring is possible. (Note: the 8th grade practice test is all right/wrong scoring.)

Illustrative Rating Form for Practice Items

Mathematics Grade 8 (continued)

Standard 8.5.2 (2000 Presidential Election)

Answered Correctly by about 67% of students[7]

Item #	Advanced	Proficient	Progressing	Beginning
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____
4	_____	_____	_____	_____
5	_____	_____	_____	_____
6	_____	_____	_____	_____

(For this exercise, do not consider the requirement for explanations for items 3 and 6.)

[7] Check only the **highest** performance level for MC items or items with "right/wrong" scoring. For tasks or items with a rubric (more points than just right/wrong scoring), check all levels for which scoring is possible. (Note: the 8th grade practice test is all right/wrong scoring.)

Appendix B Tables of results for the common districts for reading and mathematics

Tables of Results Common Districts

Reading – Grade 4 district 1-All

Tables and explanations are in the text of the report.

Reading – Grade 4 District 2-All

Table 63. Ratings of assessment for Reading standard 4.1.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	14	0	6	3	5
Central	14	0	6	4	4
Western	14	5	1	6	2
Overall	14	1.67	4.33	4.33	3.67

Table 64. Ratings of assessment for Reading standard 4.1.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	22	6	8	8	0
Central	22	1	12	9	0
Western	22	12	7	3	0
Overall	22	6.33	9	6.67	0

Table 65. Ratings of assessment for Reading standard 4.1.3 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	18	0	5	12	1
Central	18	0	7	10	1
Western	18	2	8	7	1
Overall	18	.67	6.67	9.67	1

Table 66. Ratings of assessment for Reading standard 4.1.4 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	22	3	5	9	5
Central	22	2	8	9	3
Western	22	3	7	8	4
Overall	22	2.67	6.67	8.67	4

Table 67. Ratings of assessment for Reading standard 4.1.5 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	12	1	11	0	0
Central	12	1	10	1	0
Western	12	4	7	1	0
Overall	12	2	9.33	.67	0

Table 68. Ratings of assessment for Reading standard 4.1.6 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	19	12	5	1	1
Central	19	4	14	0	1
Western	19	10	7	1	1
Overall	19	8.67	8.67	.67	1

Reading – Grade 8 District 1-All

Table 69. Ratings of assessment for Reading standard 8.1.1 across three regional meetings from District 1-All.

Region	Items*	Advanced	Proficient	Progressing	Beginning
Eastern	22	2	13	16	4
Central	22	2	13	8	4
Western	22	1	10	10	17
Overall	22	1.67	12	11.33	8.33

*Several items are scored using a rubric, thus the sum of the items may exceed the number of items in the assessment of this standard.

Table 70. Ratings of assessment for Reading standard 8.1.2 across three regional meetings from District 1-All.

Region	Items*	Advanced	Proficient	Progressing	Beginning
Eastern	15	0	7	15	0
Central	15	0	0	15	0
Western	15	0	7	4	8
Overall	15	0	4.67	11.33	2.67

*Several items are scored using a rubric, thus the sum of the items may exceed the number of items in the assessment of this standard.

Table 71. Ratings of assessment for Reading standard 8.1.3 across three regional meetings from District 1-All.

Region	Items*	Advanced	Proficient	Progressing	Beginning
Eastern	23	13	20	3	2
Central**	23	4	8	0	0
Western	23	13	15	17	17
Overall	23	10	14.33	6.67	6.33

*Several items are scored using a rubric, thus the sum of the items may exceed the number of items in the assessment of this standard.
**Teachers in the Central region did not rate all the items.

Table 72. Ratings of assessment for Reading standard 8.1.4 across three regional meetings from District 1-All.

Region	Items*	Advanced	Proficient	Progressing	Beginning
Eastern	20	1	4	16	2
Central	20	1	5	15	0
Western	20	2	5	0	15
Overall	20	1.33	4.67	10.33	5.67

*Several items are scored using a rubric, thus the sum of the items may exceed the number of items in the assessment of this standard.

Table 73. Ratings of assessment for Reading standard 8.1.5 across three regional meetings from District 1-All.

Region	Items*	Advanced	Proficient	Progressing	Beginning
Eastern	10	0	3	10	3
Central	10	0	0	7	3
Western	10	0	0	3	10
Overall	10	0	1	6.67	5.33

*Several items are scored using a rubric, thus the sum of the items may exceed the number of items in the assessment of this standard.

Table 74. Ratings of assessment for Reading standard 8.1.6 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	5	0	0	0	5
Central	5	0	0	0	5
Western	5	0	0	0	5
Overall	5	0	0	0	5

Reading – Grade 8 District 2-All

Table 75. Ratings of assessment for Reading standard 8.1.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	27	0	8	12	7
Central	27	4	6	9	8
Western	27	4	2	6	15
Overall	27	2.67	5.33	9	10

Table 76. Ratings of assessment for Reading standard 8.1.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	21	0	2	14	5
Central	21	4	2	8	7
Western	21	0	0	0	21
Overall	21	1.33	1.33	7.33	11

Table 77. Ratings of assessment for Reading standard 8.1.3 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	28	0	25	3	0
Central*	28	3	10	1	0
Western	28	5	14	4	5
Overall	28	2.67	16.33	2.67	1.67

*Teachers in the Central region did not rate all the items.

Table 78. Ratings of assessment for Reading standard 8.1.4 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	18	0	14	4	0
Central*	18	0	7	1	0
Western	18	1	11	5	1
Overall	18	.33	10.67	3.33	.33

*Teachers in the Central region did not rate all the items.

Table 79. Ratings of assessment for Reading standard 8.1.5 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	19	0	10	9	0
Central*	19	0	1	5	1
Western	19	0	9	7	3
Overall	19	0	6.67	7	1.33

*Teachers in the Central region did not rate all the items.

Table 80. Ratings of assessment for Reading standard 8.1.6 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Reading – Grade 12 District 1-All

Table 81. Ratings of assessment for Reading standard 12.1.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	11	0	0	4	7
Central	11	0	0	5	6
Western	11	0	0	9	2
Overall	11	0	0	6	5

Table 82. Ratings of assessment for Reading standard 12.1.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	8	1	1	6	3
Central	8	1	5	1	1
Western	8	0	0	4	4
Overall	8	.67	2	3.67	2.67

*One assessment task was a 10-item checklist that was treated as a single task. Only the teachers in the Eastern regional meeting rated this task as though it is rubric scored.

Table 83. Ratings of assessment for Reading standard 12.1.3 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	16	0	5	11	0
Central	16	0	4	6	6
Western	16	0	5	10	1
Overall	16	0	4.67	9	2.33

Table 84. Ratings of assessment for Reading standard 12.1.4 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	8	1	7	2	1
Central	8	1	4	2	1
Western	8	2	2	3	1
Overall	8	1.33	4.33	2.33	1

*The teachers in the Eastern region treated one assessment task as a rubric scored item.

Table 85. Ratings of assessment for Reading standard 12.1.5 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	23	0	8	14	1
Central*	23	1	3	3	3
Western	23	0	3	20	0
Overall	23	.33	4.67	12.33	1.33

*Teachers in the Central region did not rate all the items.

Table 86. Ratings of assessment for Reading standard 12.1.6 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	23	0	0	11	12
Central	23	0	0	5	18
Western	23	0	0	13	10
Overall	23	0	0	9.67	13.33

Reading – Grade 12 District 2-All

Table 87. Ratings of assessment for Reading standard 12.1.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	22	3	3	14	2
Central	22	0	7	6	9
Western	22	0	2	10	10
Overall	22	1	4	10	7

Table 88. Ratings of assessment for Reading standard 12.1.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	18	0	3	10	5
Central	18	0	0	0	18
Western	18	0	1	10	7
Overall	18	0	1.33	6.67	10

Table 89. Ratings of assessment for Reading standard 12.1.3 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	50	7	38	5	0
Central	50	4	46	0	0
Western	50	4	25	12	9
Overall	50	5	36.33	5.67	3

Table 90. Ratings of assessment for Reading standard 12.1.4 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	22	0	11	11	0
Central	22	0	3	3	16
Western	22	0	6	13	3
Overall	22	0	6.67	9	6.33

Table 91. Ratings of assessment for Reading standard 12.1.5 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	24	0	10	14	0
Central	24	0	5	16	3
Western	24	0	3	20	1
Overall	24	0	6	16.67	1.33

Table 92. Ratings of assessment for Reading standard 12.1.6 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	43	0	10	29	4
Central	43	0	5	6	32
Western	43	2	6	31	4
Overall	43	.67	7	22	13.33

Mathematics – Grade 4 District 1-All

Table 93. Ratings of assessment for Mathematics standard 4.1.5 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Table 94. Ratings of assessment for Mathematics standard 4.2.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	47	13	27	7	0
Central	47	22	22	3	0
Western	47	10	34	3	0
Overall	47	15	27.67	4.33	0

Table 95. Ratings of assessment for Mathematics standard 4.3.4 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	4	0	2	2	0
Central	4	0	2	2	0
Western	4	0	2	2	0
Overall	4	0	2	2	0

Table 96. Ratings of assessment for Mathematics standard 4.4.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	10	5	0	3	2
Central	10	5	1	2	2
Western	10	1	4	3	2
Overall	10	3.67	1.67	2.67	2

Table 97. Ratings of assessment for Mathematics standard 4.5.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	7	1	3	2	1
Central	7	1	1	5	0
Western	7	1	3	3	0
Overall	7	1	2.33	3.33	.33

Table 98. Ratings of assessment for Mathematics standard 4.6.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	17	6	8	1	2
Central	17	3	14	0	0
Western	17	7	9	1	0
Overall	17	5.33	10.33	.67	.67

Mathematics – Grade 4 District 2-All

Table 99. Ratings of assessment for Mathematics standard 4.1.5 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	5	5	5	5	5
Central	5	3	2	0	0
Western	5	0	5	0	0
Overall	5	2.67	4.	1.67	1.67

*The teachers in the Eastern region treated all assessment tasks as rubric-scored items.

Table 100. Ratings of assessment for Mathematics standard 4.2.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	28	22	6	0	0
Central	28	17	10	1	0
Western	28	17	6	5	0
Overall	28	18.67	7.33	2	0

Table 101. Ratings of assessment for Mathematics standard 4.3.4 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	4	2	2	0	0
Central	4	0	4	0	0
Western	4	0	1	3	0
Overall	4	.67	2.33	1	0

Table 102. Ratings of assessment for Mathematics standard 4.4.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	17	6	3	8	0
Central	17	2	11	2	2
Western	17	0	13	2	2
Overall	17	2.67	9	4	1.33

Table 103. Ratings of assessment for Mathematics standard 4.5.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	25	4	15	10	5
Central	25	6	15	3	1
Western	25	3	16	6	0
Overall	25	4.33	15.33	6.33	2

*The teachers in the Eastern region treated five assessment tasks as rubric-scored items.

Table 104. Ratings of assessment for Mathematics standard 4.6.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Mathematics – Grade 8 District 1-All

Table 105. Ratings of assessment for Mathematics standard 8.1.4 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	26	0	4	26	0
Central*	26	0	9	23	0
Western	26	0	4	22	0
Overall	26	0	5.67	23.67	0

*The teachers in both the Eastern region and the Central region treated some assessment tasks as rubric-scored items.

Table 106. Ratings of assessment for Mathematics standard 8.2.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	16	0	8	8	0
Central	16	0	1	15	0
Western	16	0	1	4	11
Overall	16	0	3.33	9	3.67

Table 107. Ratings of assessment for Mathematics standard 8.3.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	20	1	7	12	0
Central	20	1	8	11	0
Western	20	0	6	14	0
Overall	20	.67	7	12.33	0

Table 108. Ratings of assessment for Mathematics standard 8.4.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	40	0	3	30	7
Central	40	0	4	26	10
Western	40	0	3	30	6
Overall	40	0	3.67	28.67	7.67

Table 109. Ratings of assessment for Mathematics standard 8.5.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	7	1	2	1	3
Central	7	1	2	2	2
Western	7	0	3	3	1
Overall	7	.67	2.33	2	2

Table 110. Ratings of assessment for Mathematics standard 8.6.3 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern*	3	0	3	3	0
Central	3	0	3	0	0
Western	3	3	0	0	0
Overall	3	1	2	1	0

*The teachers in the Eastern region treated all three assessment tasks as rubric-scored items.

Mathematics – Grade 8 District 2-All

These results are reported in the body of the report.

Mathematics – Grade 12 District 1-All

Table 111. Ratings of assessment for Mathematics standard 12.1.2 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	17	0	1	7	9
Central	17	0	1	16	0
Western	17	0	2	13	2
Overall	17	0	1.33	12	3.67

Table 112. Ratings of assessment for Mathematics standard 12.2.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	15	1	11	3	0
Central	15	1	10	4	0
Western	15	8	4	3	0
Overall	15	3.33	8.33	3.33	0

Table 113. Ratings of assessment for Mathematics standard 12.3.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	12	0	0	10	2
Central	12	0	1	8	3
Western	12	0	3	6	3
Overall	12	0	1.3	8	2.67

Table 114. Ratings of assessment for Mathematics standard 12.4.5 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	11	2	6	3	0
Central	11	2	6	3	0
Western	11	2	6	3	0
Overall	11	2	6	3	0

Table 115. Ratings of assessment for Mathematics standard 12.5.1 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	9	0	2	2	0
Central	9	4	2	2	1
Western	9	2	4	1	2
Overall	9	2	2.67	1.67	1

Table 116. Ratings of assessment for Mathematics standard 12.6.3 across three regional meetings from District 1-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	16	2	13	1	0
Central	16	3	13	0	0
Western	16	0	16	0	0
Overall	16	1.67	14	.33	0

Mathematics – Grade 12 District 2-All

Table 117. Ratings of assessment for Mathematics standard 12.1.2 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	14	14	0	0	0
Central		10	4	0	0
Western	14	8	6	0	0
Overall	14	10.67	3.33	0	0

Table 118. Ratings of assessment for Mathematics standard 12.2.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Table 119. Ratings of assessment for Mathematics standard 12.3.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Table 120. Ratings of assessment for Mathematics standard 12.4.5 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	34	0	12	20	2
Central	34	0	17	15	2
Western	34	7	10	17	0
Overall	34	2.33	13	17.33	1.33

Table 121. Ratings of assessment for Mathematics standard 12.5.1 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	NR				
Central	NR				
Western	NR				
Overall	NR				

Table 122. Ratings of assessment for Mathematics standard 12.6.3 across three regional meetings from District 2-All.

Region	Items	Advanced	Proficient	Progressing	Beginning
Eastern	11	2	9	0	0
Central	11	2	8	1	0
Western	11	3	8	0	0
Overall	11	2.33	8.33	.33	0

Appendix C Illustrative Evaluation form

EVALUATION

Buros Test Sufficiency Study - Reading

Lincoln, 2003

The purpose of this evaluation is to learn your reactions to and perceptions of the various components of the NRT/CRT Sufficiency Study. Please answer each question honestly and accurately; it is very important that we have your reactions to the activities of this study. Please do not put your name on this evaluation form, as we want your responses to be anonymous. Thank you for your time is completing this evaluation.

Part 1: Orientation

The orientation consisted of several components: Overview of the study, Overview of Performance Descriptors, Overview of NAEP Performance Levels, Description of your role.

1. Using the following scale, rate the success of each training component:

Rating of Training Success

Training Components	**Very Unsuccessful**					**Very Successful**
a. Overview of study	1	2	3	4	5	6
b. Overview of Performance Desc.	1	2	3	4	5	6
d. Description of your role	1	2	3	4	5	6

2. How would you rate the amount of time allocated to the Orientation?

   a. Too much time was allocated to orientation.
   b. The right amount of time was allocated to orientation.
   c. Too little time was allocated to orientation.

3. How would you rate the amount of time allocated to the practice test?

   a. Too much time was allocated to practice test.
   b. The right amount of time was allocated to practice test n.
   c. Too little time was allocated to o practice test.

Part 2: Rating the NRT Reading questions at your grade level.

4. How comfortable do you feel about your ratings for NRT Reading questions at your grade level?

    a. Comfortable
    b. Somewhat Comfortable
    c. Not Very Comfortable
    d. Not at all Comfortable

5. How did you feel about the time allocated to rating the NRT Reading questions at your grade level?

    a. More than enough time was allotted to complete these ratings.
    b. There was sufficient time to complete the ratings.
    c. There was just barely enough time to complete the ratings s.
    d. More time needed to be allotted to complete these ratings.

Part 3. Rating the CRT Reading questions at your grade level.

6. How comfortable do you feel about your ratings for CRT Reading questions at your grade level?

    a. Comfortable
    b. Somewhat Comfortable
    c. Not Very Comfortable
    d. Not at all Comfortable

7. How did you feel about the time allocated to rating the CRT Reading questions at your grade level?

    a. More than enough time was allotted to complete these ratings.
    b. There was sufficient time to complete the ratings.
    c. There was just barely enough time to complete the ratings s.
    d. More time needed to be allotted to complete these ratings.

Part 4: Overall Evaluation of the study

8. Overall, how would you rate the success of this study?

    a. Totally Successful
    b. Successful
    c. Unsuccessful
    d. Totally Unsuccessful

9. How would you rate the organization of this study?

   a. Totally Successful
   b. Successful
   c. Unsuccessful
   d. Totally Unsuccessful

10. Please provide any comments you feel would be helpful to us in planning future studies.

Thank you for your involvement in this study!

Comments from teachers on the evaluation form

Evaluation Comments Lincoln

Grade 4 Reading:

The info about the place where the study was to be held and exactly what we were going to be doing could have been more specific. I did like this process and I felt that it was very good for my professional development.

I felt that this experience has provided me with some necessary skills that I could use for my own purpose and to share with others.

Had difficulty matching items with rubrics (on some items).

It is always nice to be able to share thoughts and ideas with other teachers in your own grade level and see other CRT tests.

Good preparation for the assigned task. Leader was personable, knowledgeable and very helpful. It was a great experience!!!!

Grade 8 Reading:

Time allocated seemed about right, but maybe working 3 half days instead of a day and a half would give a better evaluation of the questions.

Great!

Grade 12 Reading:

It was an interesting experiment!

More time for high school evaluation – warmer climate!

I would have liked more information (where and when I was to appear to participate), but that could be a problem with the communication in my district.

Please have columns on description page match those on the rating sheets. Lose the flies. Turn down the A/C or let us build a fire.

Fly control was at the beginning level ☺

Grade 4 Math:

District is paid for subs, but a stipend is necessary to continue in this process, people/teachers are burning out.

Taught me so much! Some compensation even though sub pay for district. I worked non-contract hours to get ready to have a sub! (Just a thought).

I felt frustrated at times – especially when had to share materials because there weren't enough individual copies.
Use assessments that are well-done (CRT) for practice examples.
Align practice items recording sheet going the same direction as definitions of student performance sheets for less confusion (e.g., beginning, progressing, proficient, advanced).
Offer stipend for participation – although I was happy to assist – 4th, 8th, and 11th grade teachers are constantly being asked to be out of our classrooms to work on standards and it is an added burden to get ready for a substitute. Even if it is a paid day, I was out at school until 9 PM getting ready.

This process would have been more efficient with a better defined rubric.

This was a Very good learning experience – very worthwhile!

Grade 8 Math:

None

Grade 12 Math:

This provided good feedback for our own district.

I was enlightened.

Evaluation Comments Kearney

Grade 4 Reading:

Color code the census form.

Grade 8 Reading:

I'm not sure how useful this information is because so many of the assessments did not assess the standard they were supposed to be assessing.

Grade 12 Reading:

I would double-check the alignment of the assessments to make sure they correspond to the proper standards. This (the mis-alignment) was a frustration for our group.

I feel our work would have been more successful if the assessments had been aligned to the standards prior to our rating. The information would have been quite a bit more useful.

It has been quite interesting to me to review and evaluate the assessments used by our schools. I was amazed at how similar the assessments were. I'm curious about the sources of each assessment (Internet, ESUs, etc.). How many were developed by the local schools? Were they duplicated and then sent out to other schools? I enjoyed the process of evaluation and discussion of each question on the assessments. It's always intriguing to hear other teacher's justification for a response. I thought Jim, Chad, & Renee did a great job with the workshop and I fount that their instructions were very explicit. It was well organized.

I wonder if enough school's assessments were considered to compose a valid study.

I learned a lot to take back to my district to evaluate our own assessments.

Excellent organization. Thanks Chad for taking time to answer our questions.

Grade 4 Math:

Better communication is my only suggestion!

Communication prior to testing.

Gave me some great insight to look for in my own school's assessments.

Grade 8 Math:

Bring a box with hands on things (silly putty, play doh…) for teachers to "play" with.

This was an excellent experience for my growth as an educator. Very positive, give confidence as a teacher.

Grade 12 Math:

Enjoyed the process.

I enjoyed the process and feel like I gained knowledge.

Going through the practice set was very helpful. The rubric was very helpful as well.

Evaluation Comments Scottsbluff

Grade 4 Reading:

This was a very well organized study. The directions were clear and the purpose became even more clear as we got into this. I have been really frustrated with the lack of direction and explanation about tests up to this point and this really helped. I feel that while the purpose was to help with the study, I really benefited from this as well. It was a real eye opener and the collaboration was great! Thank you!

Good study! I really hope that we can effectively use all these workshops to build on the standards of Nebraska. I still feel that a good start would be to develop a State curriculum for all subjects. Too many students moving around state not meeting standards.

The administration did not tell us what we were going to do. I believed we were reflecting on our districts assessments to improve communication. Thank you, this was a valuable experience.

This study provided me with ideas to take back and put to use in my district. I thought the process was practical and allowed for us to realize how our own assessments and scoring practices can be improved.

Interesting. Thought provoking.

Grade 8 Reading:

We learned how others view the definitions and how to clarify them. This study provides us time to meet with our peers and come to a common ground to help our students understand these standards and how we, as teachers, can help each student become aware of specific reading behaviors.

I've participated twice, Lincoln and Scottsbluff. I've enjoyed the work BUT, I am not at all confident that our ratings are reliable. Group dynamics affect the consensus too much. In Lincoln we had a "battleaxe:" and her opinion won out because people were too tired or afraid to argue with her. In Scottsbluff we had a very knowledgeable group but very congenial and eager to reach consensus. We "met in the middle" a fair number of times. The benefits of participation was a thorough discussion of standards and items by the group. I think the ratings will vary by group. I'd like to know the results. I do enjoy working for Buros because I learn so much. I rated the overall evaluation unsuccessful. Define success. If it was to reliably classify items I don't think our results are reliable.

The time frame was better than other evaluations I've done. We had plenty to do, but we had the right amount of time. (We'd gotten done quite early on other evaluations I've done).

This was an awesome opportunity to view other tests, think about what we are doing, and see where we may need to improve.

Grade 12 Reading:

Excellent staff.

The green sheets used as a "rubric" were inadequate. We relied heavily on classroom experience.

Very well directed. I enjoyed the two days very much.

Chad was such a good listener, helpful with questions. Indeed professional. Thanks Chad for clarifying so many uncertainties we had.

Perhaps 2 full days would have been better, especially since validity is the main issue. This became mind-numbing with the switches in complexity and format changes.

I hope that what we did was based on correct criteria. I was frustrated when the standards did not relate to the questions.

Grade 4 Math:

Learned a lot especially about using rubrics in my classroom. Thanks!

Definitions need to be in the same order as scoring sheet. It is wonderfully to be treated professionally. Thank you.

It is always fun to interact with teaching colleagues and with the Buros Center. I always learn and feel my own understanding of testing, standards and validation has grown.

Grade 8 Math:

Great planning and great food!

I think the groups should just discuss each one as they go instead of doing individually then as a group.

Grade 12 Math:

The descriptors sheet and the rating sheets need to have the levels in the same order.

Matching questions to definitions of the different levels was very helpful to me personally. Makes it easier to look at our own assessments critically.